Crystallographic studies of central metabolic enzymes from the hyperthermophilic Archaea Pyrococcus woesei and Sulfolobus solfataricus

Bell, Graeme S.

Award date:
1999

Awarding institution:
University of Bath

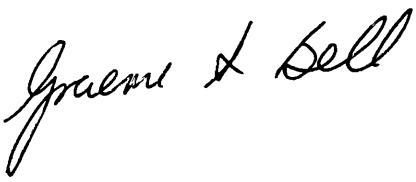
Crystallographic Studies of Central Metabolic Enzymes from the Hyperthermophilic Archaea *Pyrococcus woesei* and *Sulfolobus solfataricus*

submitted by Graeme S. Bell

for the degree of PhD

University of Bath 1999

COPYRIGHT



Graeme Bell

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

In an attempt to further our understanding of the structural basis of protein hyperthermostability, we have chosen to study crystal structures of central metabolic enzymes from the Archaea. The gene encoding for triosephosphate isomerase (TIM) from the hyperthermophilic Archaeon *Pyrococcus woesei* has been cloned and sequenced, and the enzyme overexpressed in *Escherichia coli*[1]. The recombinant protein was subsequently purified to homogeneity and co-crystallised with two substrate analogues, yielding crystals suitable for X-ray diffraction studies. Predominantly two crystal forms were obtained; orthorhombic ($P2_12_12$) and monoclinic ($P2_1$). Complete data sets have been collected for both crystal forms to a resolution of 4.0Å and 2.6Å respectively. The available crystal structures of TIM from several Eukaryal and Bacterial sources were used in extensive efforts to solve the structure by the Molecular Replacement method. These were unsuccessful, despite the additional application of the knowledge of non-crystallographic symmetry. An attempt was therefore made to solve the structure using the heavy atom Isomorphous Replacement method. A single heavy atom protein derivative was obtained at a later stage of the project although initial attempts to elucidate the crystal structure using the Single Isomorphous Replacement phases combined with non-crystallographic symmetry averaging have not allowed determination of the structure. In the final year of this work, a second project was embarked upon due to the availability of a crystal of citrate synthase from another hyperthermophilic Archaeon; *Sulfolobus solfataricus*. Data were collected for a monoclinic ($P2_1$) crystal form to a resolution of 2.7Å, and the structure was solved via Molecular Replacement, using the *Thermoplasma acidophilum* citrate synthase crystal structure as a phasing model. Restrained refinement of the model using the Maximum Likelihood method, gave a final R-factor of 20.8% (R_{free}=28.5%). This model was used for a comparative study with the several citrate synthase crystal structures already available from organisms growing at a wide range of temperatures. Results from this work have contributed to the ongoing research into the molecular basis of protein thermoadaption; highlighting structural features correlating with protein thermostabilty in general, and specifically emphasising the importance of

[1] by Michael Kohlhoff at the University of Essen, Germany.

oligomer interface contacts and electrostatic interactions as a means of stabilisation in thermophilic citrate synthases.

ACKNOWLEDGEMENTS

I would like to thank my main supervisor Garry Taylor, for all his guidance and patience over the last three years, and to whom I could have asked no more in his role. Michael Danson and David Hough have been of great help, and I thank them both for their enthusiasm and biochemical enlightenment. I would also like to thank all research staff and postgraduate students who have contributed to the project and in particular for the time, advice and expertise of Rupert Russell, whose work was inextricably linked to my own. Thankyou also to Peter Oflaherty for help with the endless technical problems during purification. On the TIM project, I owe thanks to Reinhard Hensel and Michael Kohlhoff for initiating the work and continuing support throughout. For the latter work I also thank Helen Lawrence for her contribution and Professor Steve Homans at the University of St. Andrews for carrying out the NMR experiment. The BBSRC are also thanked for funding.

Finally I owe special thanks to Mum and Dad for their constant support and to Caroline for putting up with me, for all the sacrifices made, and for picking me up during the most difficult times.

ABBREVIATIONS

Amino Acids (one and three letter codes)

Alanine	ALA	A	Leucine	LEU	L
Arginine	ARG	R	Lysine	LYS	K
Asparagine	ASN	N	Methionine	MET	M
Aspartate	ASP	D	Phenylalanine	PHE	F
Cysteine	CYS	C	Proline	PRO	P
Glutamine	GLN	Q	Serine	SER	S
Glutamate	GLU	E	Threonine	THR	T
Glycine	GLY	G	Tryptophan	TRP	W
Histidine	HIS	H	Tyrosine	TYR	Y
Isoleucine	ILE	I	Valine	VAL	V

a,b,c and α,β,γ	real space cell axes lengths and inter-axial angles
α	phase angle
α,β,γ	Eulerian angles
ATP	adenosine triphosphate
B-factor (or B)	isotropic temperature factor
1,3-BPG	1,3 bisphosphoglycerate
BSA	bovine serum albumin
CS	citrate synthase (E.C. 4.1.3.7)
CoA	coenzyme A
EDTA	(disodium) ethylenediamine tetraacetate
Da	Daltons
DHAP	dihydroxyacetone phosphate
f_j	atomic scattering factor
F_c, $\lvert F_c \rvert$	calculated structure factor and its amplitude
F_o, $\lvert F_o \rvert$	observed sructure factor and its amplitude
FOM	figure of merit
GAP	glyceraldehyde-3-phosphate

GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GLUDH	glutamate dehydrogenase
HEPES	(N-[2-hydroxyethyl]piperazine-N'-[2-ethanesulphonic acid])
IMPDH	Isopropylmalate dehydrogenase
IPTG	isopropyl β-D-thiogalactopyranoside
kDa	kilodaltons
K_M	Michaelis constant
MES	(2-[N-morpholino]ethanesulphonic acid)
M.W.	molecular weight
NADH	nicotinamide-adenine dinucleotide
NADPH	nicotinamide-adenine dinucleotide phosphate
NCS	non-crystallographic symmetry
OAA	oxaloacetate
PAGE	polyacrylamide gel electrophoresis
PCMB	para-chloromercuribenzoate
PCMBS	para-chloromercuribenzene sulfonate
PDB	protein data bank
PEG	polyethylene glycol
PGK	phosphoglycerate kinase
ϕ,ψ,κ or ω,ϕ,κ	spherical polar angles
PIP	Di-μ-iodobis(ethylene diamine)diplatinum(II) nitrate
PMSF	phenylmethanesulfonyl fluoride
R-factor (R)	Crystallographic residual
RMS	root mean square
rRNA	ribosomal ribonucleic acid
SDS	sodium dodecyl sulphate
σ	standard deviation
TIM	triosephosphate isomerase (E.C. 5.3.1.1)
TRIS	tris-(hydroxymethyl)-methylamine
V	unit cell volume
V_{MAX}	Maximal reaction velocity
x,y,z	Cartesian coordinate axes

CONTENTS

CHAPTER 1

Introduction

1.1 Overview

The overall aim of this project was to undertake structural studies on enzymes from hyperthermophilic Archaea, due to their evolutionary interest and potential use in biotechnology. In particular, enzymes from central metabolic pathways have been targeted as good model systems due to their ubiquity in nature. The core of this study was the crystal structure analysis of the central metabolic enzymes triosephosphate isomerase from *Pyrococcus woesei* (*Pw*TIM) and citrate synthase from *Sulfolobus solfataricus* (*Ss*CS). *P. woesei* is a strictly anaerobic, heterotrophic Archaeon, originally isolated from sediment samples obtained from marine solfataric vents [Zillig *et al*, 1987] and grows optimally at a temperature of approximately 100°C and neutral pH. *S. solfataricus* was isolated from the low pH environment of hot mud pools near Naples [De Rosa *et al*, 1975] and is a facultative autotroph, exhibiting an optimum growth temperature of 87°C. Due to their high growth temperatures, these Archaea are classed as hyperthermophilic, and as such, contain enzymes adapted to extreme temperatures. Elucidation of the crystal structures of enzymes from these hosts should therefore further our knowledge of Archaeal proteins, and highlight structural features which could confer protein hyperthermostability.

1.2 The Archaea

The discovery of the Archaebacteria, reported in 1977 [Woese & Fox, 1977] was the result of initial studies based on the sequences of 16s/18s rRNA (reviewed [Woese, 1987, Olsen & Woese, 1993]) enabling classification of these micro-organisms via comparisons at the molecular level. This led to the proposal of division of organisms into three domains of life, namely Eukarya, Eubacteria and Archaea, [Woese *et al*, 1990] with the acceptance that the

Archaea were a phylogenetically distinct group with respect to the other two domains (see figure 1.1). The evolutionary relationships between these three domains has been discussed extensively [Woese, 1993, Doolittle, 1995, Keeling & Doolittle, 1995, Brown & Doolittle, 1997], with Archaea being morphologically more similar to Bacteria, yet phylogenetically more closely related to the Eukarya [Brown & Doolittle, 1997]. The phylogenetic relationship seen in figure 1.1 holds for both rRNA and 'informational molecules' within the cell. However, indications of the occurrence of horizontal gene transfer throughout evolution probably explain the more conflicting picture observed for other proteins involved in central metabolism and cell pathways [Olendzenski & Gogarten, 1998].

The Archaea span a broad range of phenotypes, existing in extreme environments; including the halophiles (growing in salt concentrations approaching saturation), acidophiles (growing in pH environments as low as pH 0), those adapted to extremes of temperature, being classed as psycrophiles (growing optimally in temperatures <10°C), thermophiles and hyperthermophiles, (the latter exhibiting optimum growth temperatures above 80-85°C) and others such as those of a barophilic phenotype (growing in extremes of pressure, typically in deep sea environments).

The Archaea are of interest for phylogenetic reasons, being thought to be one of the most primitive forms of life, but also, because of the extreme environments which they inhabit, the study of proteins from these organisms can give insight into the structural basis of adaption to extremes of temperature, salinity, pH and pressure. Thermostable proteins can be put to use in biotechnological applications [Danson *et al*, 1991, Cowan, 1992, Herbert, 1992, Adams *et al*, 1995] with examples such as DNA polymerases in the polymerase chain reaction (PCR), a wide range of biocatalysts, and proteases in washing powder. Although this topic has been well discussed, it is likely however, that the full potential of these applications has yet to be utilised.

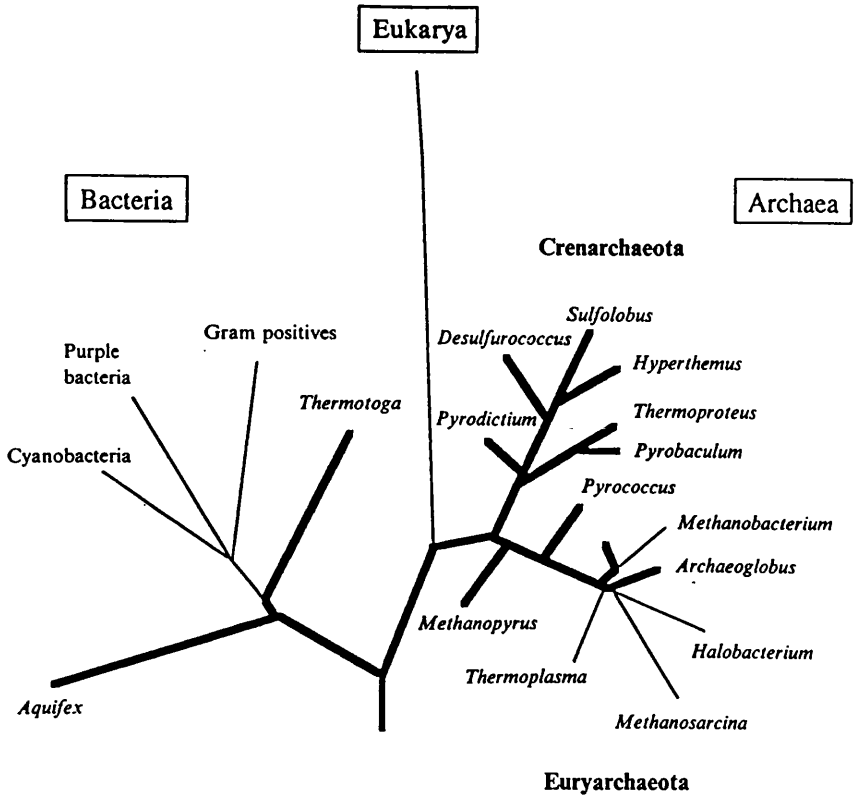


Figure 1.1: 16s based rRNA rooted universal phylogenetic tree showing the three domains [Schonheit & Schafer, 1995] and specifically those lineage's defined as hyperthermophilic (bold lines). Only two of these lineage's are from the Bacterial domain; *Aquifex* and *Thermotoga*.

1.3 Biological Context

The enzymes of central metabolism are of particular interest because their ubiquity in nature makes them excellent for comparative study. There are however several differences between central metabolism in the Archaea, and that of Eukaryal and Bacterial organisms. This has been reviewed concerning the different phenotypes mentioned above, [Danson *et al*, 1989, 1991 & 1993] and metabolism specifically in hyperthermophiles has also been more recently reviewed [Schonheit & Schafer, 1995].

The classical Embden-Meyerhof glycolytic pathway is that seen in eukaryotic cells and many anaerobic Bacteria. The enzyme phosphofructokinase (which

catalyses the conversion of fructose 6-phosphate to fructose 1,6-bisphosphate) is absent in many aerobic bacteria, and therefore glucose catabolism goes via an alternate route in these organisms (the Entner-Doudoroff pathway). Both these pathways are shown in figure 1.2.

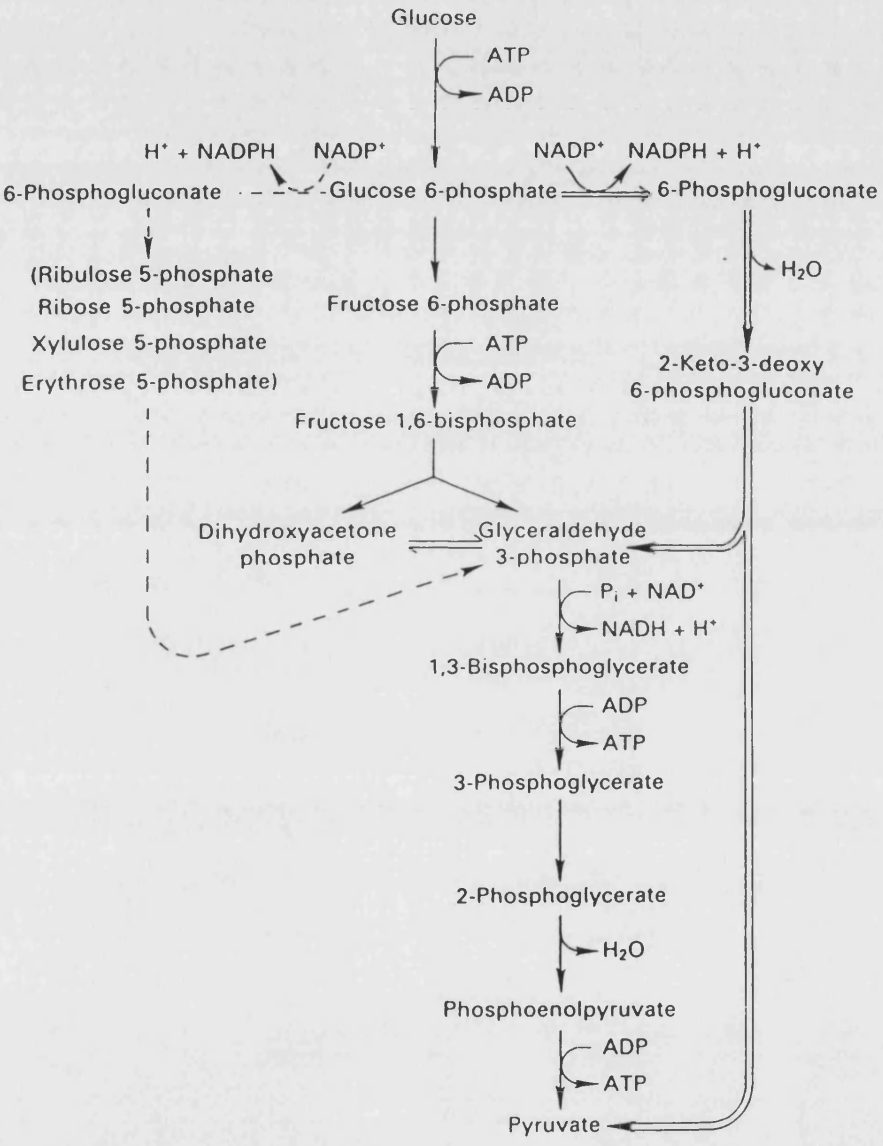


Figure 1.2: Pathways of glucose catabolism in Eukarya and Bacteria [Danson *et al*, 1991]. The classical Embden-Meyerhof glycolytic pathway, (single line), and the Entner-Doudoroff pathway (double line). The pentose-phosphate pathway which metabolises glucose for the generation of NADPH and pentose and tetrose sugars is also shown (dotted line).

Glucose is catabolised slightly differently amongst archaeal organisms. Many halophiles, such as several of the *Halobacterium* species [Tomlinson *et al*, 1974], have a modified Entner-Doudoroff pathway. It has also been demonstrated that the thermoacidophiles *Thermoplasma acidophilum* [Danson, 1989, Budgeon & Danson, 1986] and *Sulfolobus solfataricus* [De Rosa *et al*, 1984] catabolise glucose via a non-phosphorylated Entner-Doudoroff pathway. These Archaeal pathways can be seen in figure 1.3.

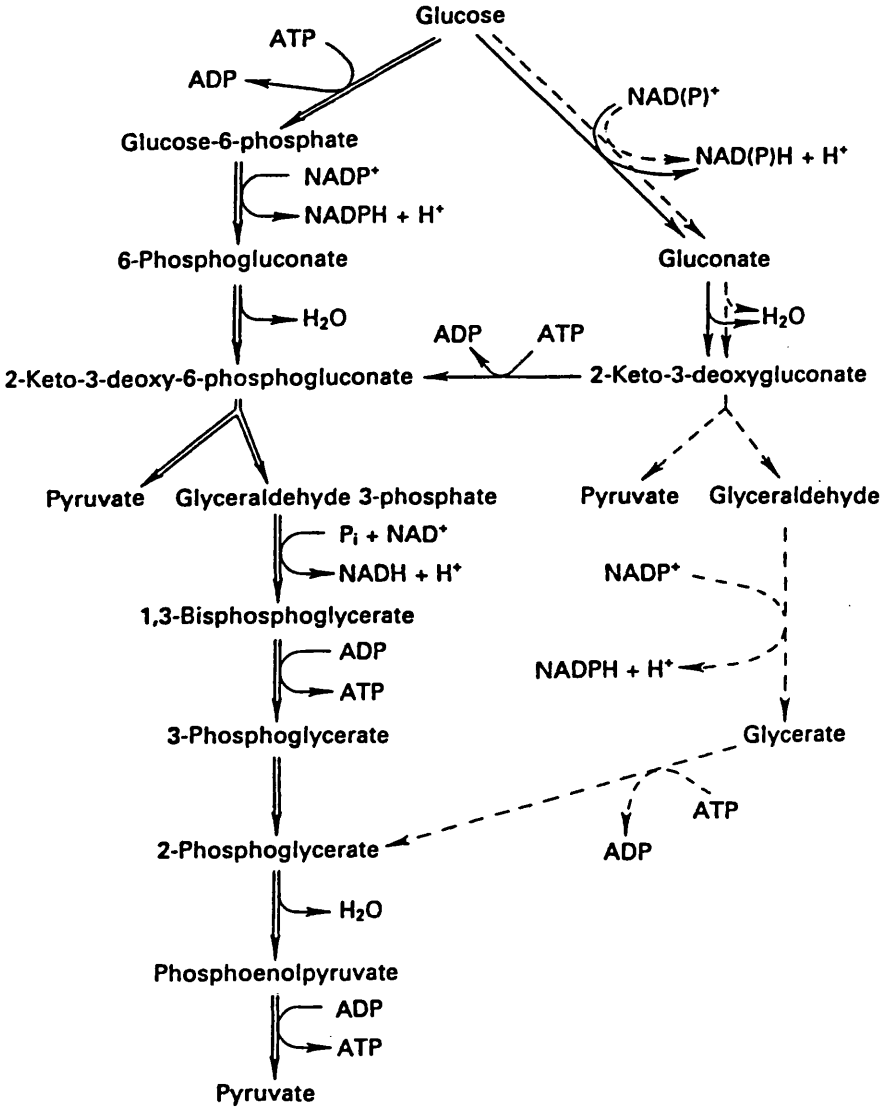


Figure 1.3: Pathways of glucose metabolism found in Archaea [Danson *et al*, 1991]. The modified Entner- Doudoroff pathway as seen in some halophiles (single line) and the non-phosphorylated Entner-Doudoroff pathway observed in the acidophiles (dotted line). The Entner-Doudoroff pathway is also shown here (double line).

Studies on *Pyrococcus furiosus*, [Schafer & Schonheit, 1993] have revealed that in addition to enzymes of the non-phosphorylated Entner-Doudoroff pathway of the thermoacidophiles, evidence was found for the existence of a modified Embden-Meyerhof pathway (with the enzyme phosphofructokinase being dependent on pyrophosphate instead of ATP) for glucose catabolism. It has also been demonstrated that although enzymes are present for both pathways, glucose catabolism only appears to proceed via the modified Embden-Meyerhoff route [Selig *et al*, 1997] (figure 1.4). It should be noted that there is no net yield of ATP in glucose catabolism via either the modified Embden-Meyerhof pathway of *P. furiosus* or the non-phosphorylated Entner-Doudoroff pathway of *S. solfataricus*.

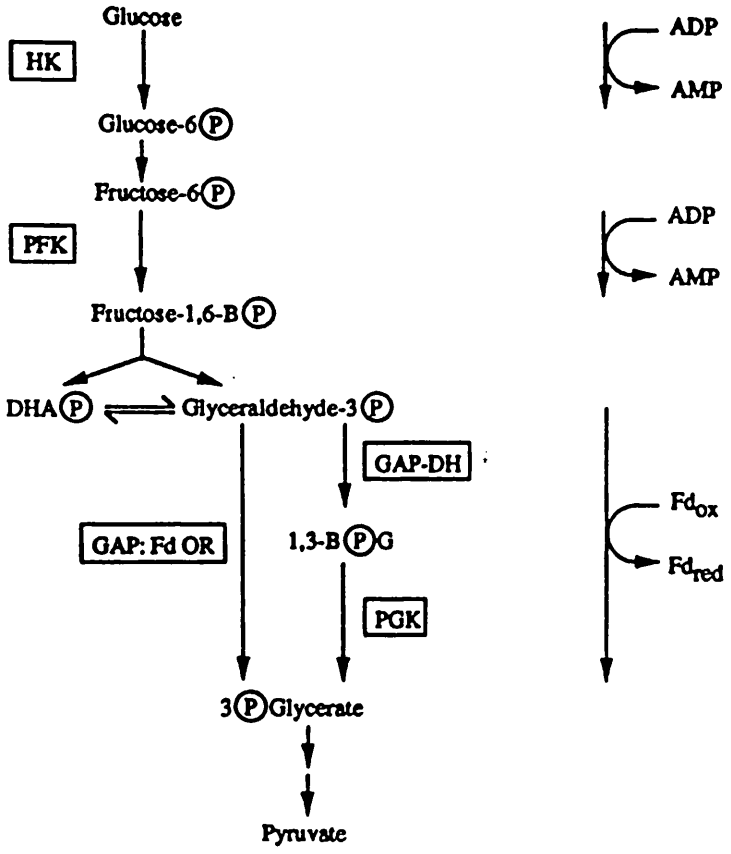


Figure 1.4: The Modified Embden-Meyerhof pathway present in *P. furiosus* compared with that seen in thermophilic bacteria [Schonheit & Schafer, 1995]. (HK; hexokinase, PFK; 6-phosphofructokinase, GAP:Fd OR; glyceraldehyde-3-phosphate ferredoxin oxidoreductase, PGKphophoglycerate kinase and GAPDH; glyceraldehyde-3-phosphate dehydrogenase)

Gluconeogenesis from acetyl-CoA or pyruvate has been shown to occur via the reversal of the Embden-Meyerhoff pathway in many hyperthermophiles [Schonheit & Schafer, 1995].

Pyruvate is oxidatively decarboxylated to yield acetyl-CoA via pyruvate:ferredoxin oxidoreductase in all hyperthermophiles studied to date [Schonheit & Schafer, 1995]. Acetyl-CoA then enters the citric acid cycle (figure 1.5), being the final stage in the oxidation of metabolites to produce NAD(P)H and $FADH_2$, (linked with the generation of ATP) and the starting point for many biosynthetic pathways. The citric acid cycle is operative as either an oxidative, reductive, or partial cycle in many organisms and the extent to which it is important varies depending on both organism and environment. Under aerobic conditions, it has been shown that *St...olobus* operates this pathway [Danson, 1988]. In the anaerobic *Pyrococcus furiosus* growing on sugars and peptides, it has been shown that pyruvate is fermented to acetate, after initial conversion to acetyl-CoA [Schafer & Schonheit, 1991] providing a net gain in ATP. The citric acid cycle is therefore present only as a partial cycle in this case, operating purely for biosynthesis [Schonheit & Schafer, 1995].

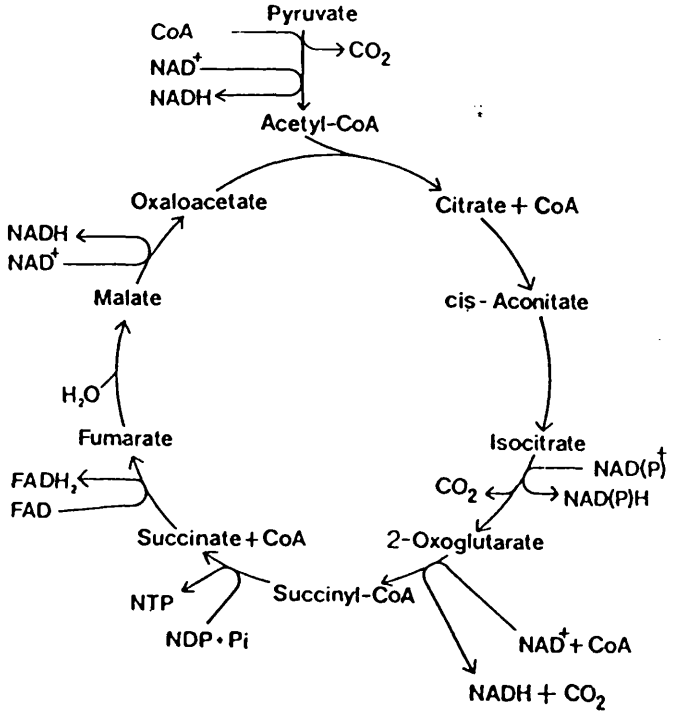
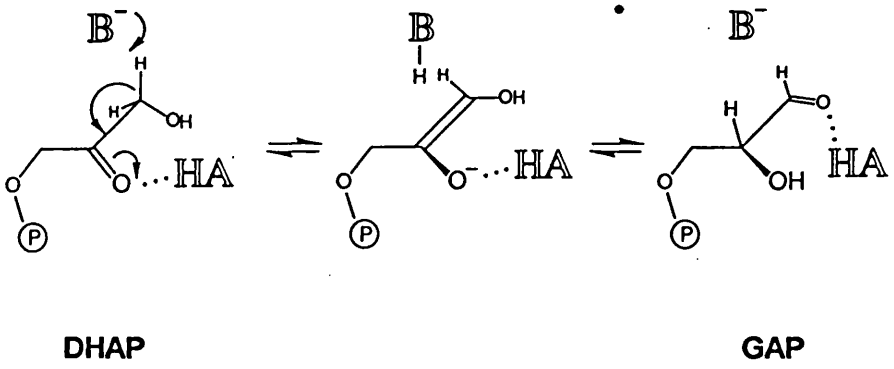


Figure 1.5: The citric acid cycle [Danson, 1993].

1.4 Triosephosphate Isomerase (TIM)

1.4.1 Metabolic Role

TIM catalyses the interconversion of D-glyceraldehyde-3-phosphate (D-GAP) and dihydroxyacetone phosphate (DHAP) in the pathways of glucose catabolism described in section 1.3. The reaction proceeds via an enediol intermediate (below). This isomerisation allows the utilisation of all carbon units generated by aldol cleavage, in the previous step of the pathway. The reaction intermediate is stabilised by a flexible protein loop, and the two enolisations are mediated by a catalytic acid and base [Knowles, 1991].



DHAP GAP

Figure 1.6: The reversible isomerisation reaction catalysed by TIM (The catalytic base (B^-) being Glu 167 and catalytic acid (HA) His 95) [Lodi & Knowles, 1991] (see also figure 1.7).

TIM represents one of the simplest and most efficient enzymes, with the rate of reaction limited by the diffusion rate, and being 10^{10} fold faster than the non-enzymatic base catalysed conversion. TIM is therefore often referred to as the 'perfect enzyme', [Knowles & Albery, 1976] however it is clear that factors such as electrostatic steering due to interaction potential in the active site, can be responsible for enhancement of these rate limits [Wade, 1996, Zhou *et al*, 1997]

TIM has also been of interest for rational drug design projects with both the *Trypanasoma brucei* [Operdoes *et al*, 1988] and malarial parasite *Plasmodium falciparum* [Velanker *et al*, 1997] glycolytic systems being studied.

1.4.2 Structure

The structure of triosephosphate isomerase is highly conserved across Bacteria and Eukarya, with a monomer of approximately 250 amino acids comprising an inner core cylinder of eight twisted parallel β-strands interconnected by eight α-helices (figures 1.8 and 1.9). All known bacterial and Eukaryal TIMs exist as homo-dimers with a total molecular weight of approximately 50-60kDa (the exception being the tetrameric fusion protein of PGK/TIM from *Thermatoga maritima* [Schurig *et al*, 1995]). Loop regions (numbered 1-8) at the face of the barrel (comprising those connecting the C-termini of the β-strands with the N-termini of the adjacent helix) contain most of the active site residues; the catalytic residues His 95 (loop 4), Glu 167 (loop 6) and the phosphate binding sites being Lys 13 (loop 1) and main chain NH groups (loops 6 and 7 and helix 8) [Lolis & Petsko, 1990, Noble *et al*, 1991, Davenport *et al*, 1991, Wierenga *et al*, 1991]. The dimer interface interaction (loops 1-4) involves a long loop (loop 3) which protrudes into a hydrophobic pocket near the active site of the other monomer [Alber *et al*, 1981] (see figure 1.9). The contrast between the structural α/β core of the TIM Barrel, with functional residues which reside within the loops, make it a good model on which to perform mutational studies, possibly altering enzymatic function, without affecting the stability of the enzyme as a whole [Borchert *et al*, 1995]. The $(\alpha/\beta)_8$ TIM barrel is also the most frequently occurring motif found in proteins [Branden, 1991], and is a common structural scaffold for enzymes which perform a diverse range of functions [Reardon & Farber, 1995]. The sequence homology of these different TIM Barrel enzymes is quite low in most cases, despite the highly conserved 3-D structure, and studies on their possibly divergent evolution has also been reviewed [Branden, 1991 ,Farber & Petsko, 1995]. There are currently eight published structures of TIM from both Eukaryal and Bacterial sources, however there is no data on the structure of Archaeal TIMs. There are six known structures of TIM isolated from mesophilic

sources; chicken muscle [Banner *et al*, 1975] , *Trypanasoma brucei* [Wierenga *et al*, 1987] , Yeast [Lolis *et al*, 1990] , *Escherichia coli* [Noble *et al*, 1993] human [Mande *et al*, 1994] and *Plasmodium falciparum* [Velanker et al, 1997]. The structure of TIM from the thermophilic bacterium *Bacillus stearothermophilus* [Delboni *et al*, 1995], (which exhibits an optimum growth temperature of 65°C) and the psychrophilic bacterium *Vibrio marinus* [Alvarez *et al*, 1998] have also been published. It has been reported that structures have been determined for *Leishmania mexicana* TIM and the thermophilic *Thermotoga maritima* TIM but are not yet in publication [Alvarez *et al*, 1998]. The authors report that a study on thermostability of TIMs is currently being carried out on these enzymes. There are also crystal structures of several mutant enzymes [Joseph-McCarthy *et al*, 1994, Borchert *et al*, 1995].

1.4.3 Reaction Mechanism

The reaction mechanism and kinetics have been extensively studied by a range of techniques in biochemistry and molecular biology, in addition to structural studies with X-ray crystallography and NMR using wild-type and mutant enzymes. Key residues in the active site are glutamate 167 acting as a base, and histidine 95 as an acid [Joseph-McCarthy *et al*, 1994, Lodi & Knowles, 1991, Komives *et al*, 1991, Davenport *et al*, 1991]. Lysine 13 has also been shown to be instrumental in the binding of substrate [Lodi *et al*, 1994] (numbering according to *T. brucei*). An additional feature of the reaction mechanism is the conformational change of the active site on the binding of substrate, concerning a major shift in a flexible eight residue loop region (loop 6) which moves 7Å between "open" and "closed" conformations, thus binding and stabilising the reaction intermediate [Joseph *et al*, 1990].

All interactions between substrate analogues and protein have been found to involve atoms of a single subunit, and therefore both active sites function independently [Wierenga *et al*, 1992]. However, it is known that TIM is only active as a dimer [Waley, 1973] and therefore studies were carried out to investigate this further by creating mutants which impede dimer formation [Mainfroid *et al*, 1996]. Mainfroid *et al* created three mutant TIMs which

interfered with the monomer-dimer equilibrium to differing degrees. The structure of an engineered monomeric TIM (via deletion of loop 3) has been determined [Borchert *et al*, 1993] and was shown to fold as a TIM barrel, however these mutational studies demonstrated the importance of dimerisation in maintaining integrity of the active site.

The similarity between the reaction mechanism of TIM with that of citrate synthase (section 1.5) have been discussed [Remington, 1992a] as both acid/base mechanisms involve the deprotonation of a carbon atom, with only the weak acids and bases of the amino-acid sidechains, and in the absence of metal ions. In the case of TIM, it is most likely that the high energy neutral enediol (and or enediolate) intermediate is generated by a concerted reaction, with His 95 acting as an acid, with Glu 167 as the base, abstracting a proton from DHAP. Surprisingly, although His 95 acts as an acid, it has been shown that the imidazole ring remains neutral over the entire pH range of activity [Lodi & Knowles, 1991] (and a similar situation is likely to be present in citrate synthase). The following isomerisation generating D-GAP is simply the reverse of this reaction (with the acid and base roles reversed). This proton 'shuttling' can be carried out with minimal movement of the catalytic sidechains [Knowles, 1991, Wierenga *et al*, 1992] (a suggested mechanism can be seen in figure 1.7).

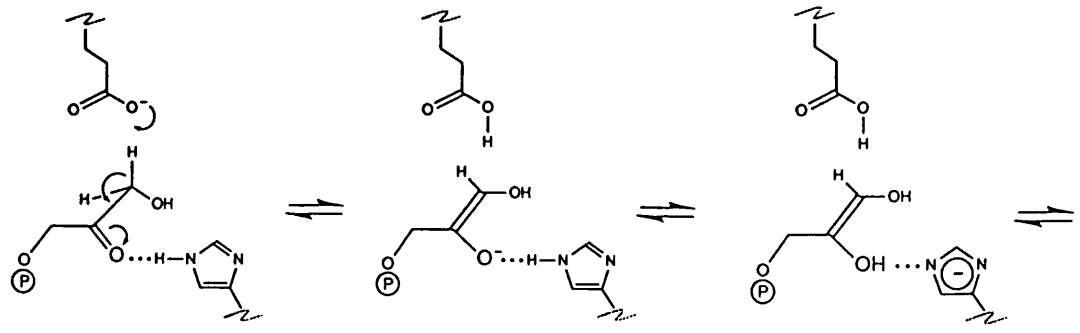


Figure 1.7: A possible mechanism for the reaction catalysed by TIM, suggesting the role of the (neutral) imidazole group of His 95 [Lodi & Knowles, 1991].

1.4.4 TIM from *P. woesei* and Homology with known TIMs

The derived protein sequence of *P. woesei* TIM comprises 224 residues [Kohlhoff *et al,* 1996] (with a monomer Mr.= 24 657 Da) and is the shortest TIM sequence known to date. In addition, ultracentrifugation and gel filtration experiments have suggested that TIM from *P. woesei* exists as a homo-tetramer. Similar experiments on TIM from *Methanothermus fervidus* (optimum growth temperature of 83°C) have also suggested that this enzyme is homo-tetrameric [Kohlhoff *et al,* 1996]. As these results are contrary to those found for Eukaryal and Bacterial TIMs (being homo-dimeric), it has been proposed that this higher state of aggregation is in fact correlated to an increase in thermostability, due to a reduction in exposed surface area and increased intersubunit contact. Also, the molecular mass of TIM fom the mesophilic methanogen *Methanobacterium bryantii* suggests that this enzyme is homo-dimeric [Reinhard Hensel, private communication], giving further evidence that the tetrameric TIM is not simply an Archaeal feature. *Pw*TIM also displays rather low homology with respect to the Bacterial and Eukaryal TIMs. Pairwise sequence identities with the 'structural' TIMs range from 20.8% (*B. stearothermophilus*) to 26.5% (*T. brucei*). In comparison, known Bacterial and Eukaryal sequences share an identity of approximately 40%. However, when *Pw*TIM is aligned with the enzyme from the hyperthermophilic Archaeon *M. Jannaschii* this gives a sequence identity of 65%, suggesting the possibility that the latter enzyme may also exist as a homotetramer. A structure based sequence alignment can be seen in figure 1.10 showing the conserved catalytic and dimer interface binding residues). Structural comparison of the TIM dimers displayed an average RMS distance of 1.5Å for 480 alpha-carbon atoms, with the major differences occurring in several loop regions. The largest apparent deletion in sequence (in both the *P. woesei* and *M. jannaschii* sequence) occurrs in the region containing a large part of helix 5 and the following loop (see figure 1.9). This region is shown to protrude into solvent in the dimeric structures, and may be involved in the formation of the tetramer. Full structural analysis should reveal some of these speculations.

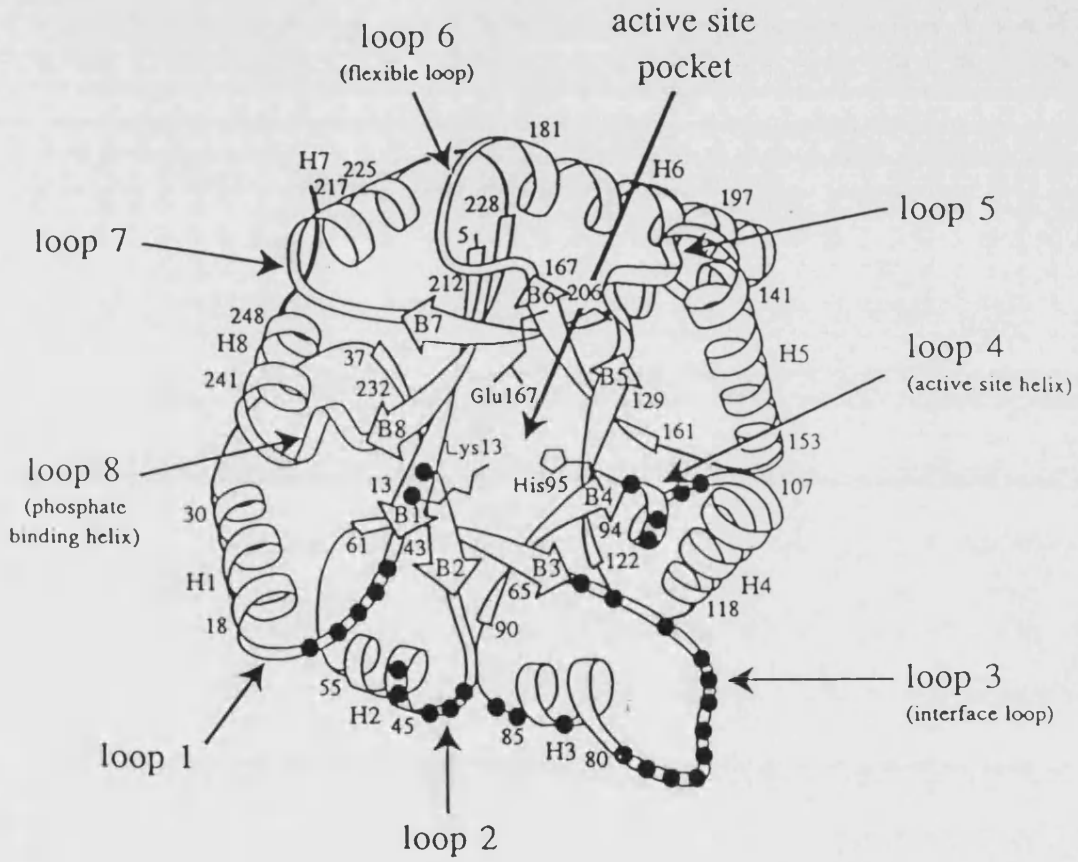


Figure 1.8: Schematic diagram of a TIM Monomer showing explicit labelling of the 8 loop regions connecting C-termini of β-strands with the N-termini of subsequent α-helices. Black spots show the location of interface residues. [Wierenga *et al*, 1992]

Figure 1.9: Ribbon diagram showing two orientations of the *E. coli* TIM homo-dimer, highlighting the location of helix 5 and the following loop (pale green) which may be involved in intersubunit contacts of the *Pw*TIM tetramer. The diagram was created using **MOLSCRIPT** [Kraulis, 1991].

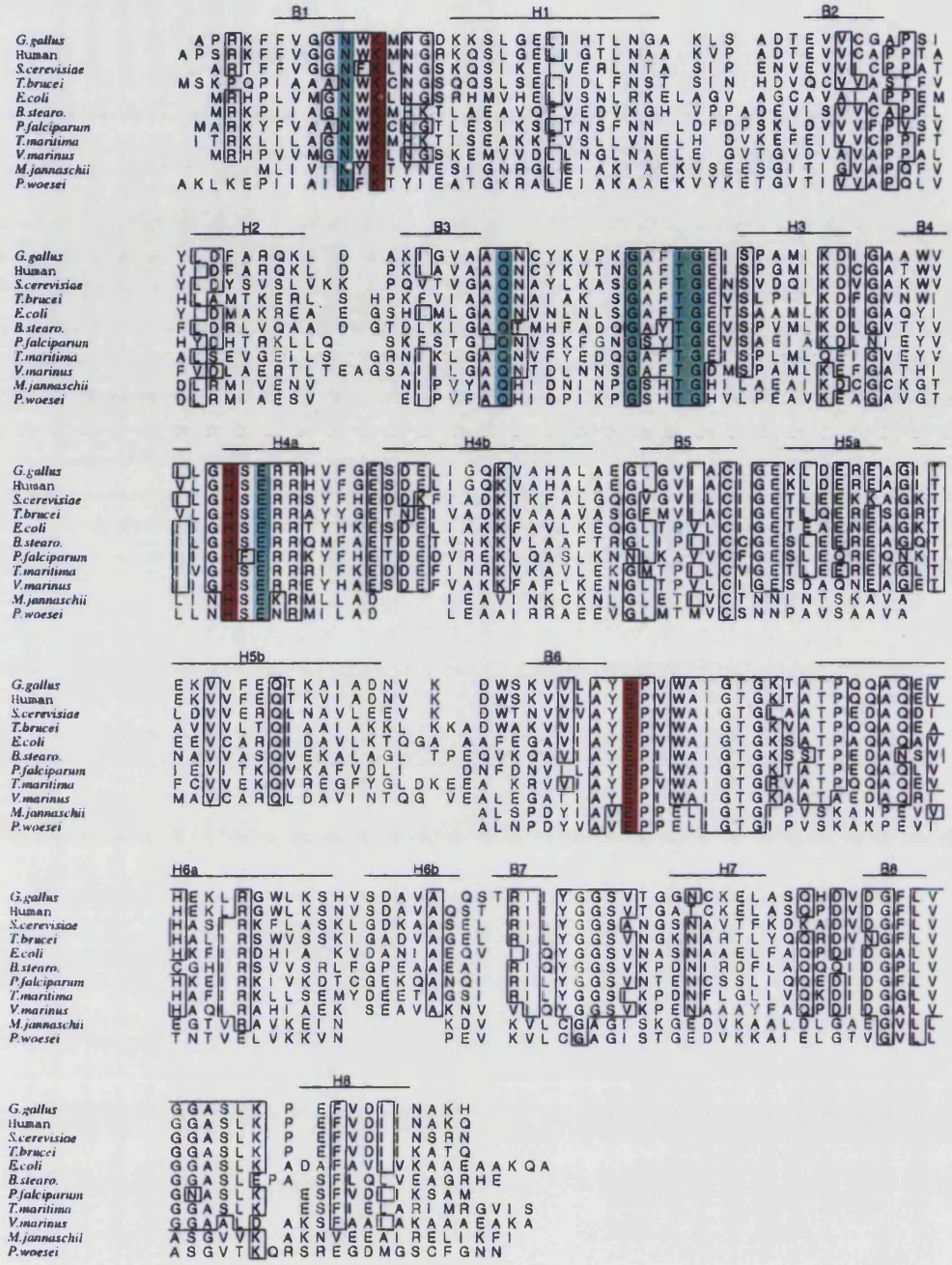


Figure 1.10: Structure based sequence alignment of *Pw*TIM with the other structural TIMs and TIM from *M jannaschii*. Identical residues are boxed, functional residues K13, H95 and E167 are shown in red and fully conserved interface residues are in green. Structural alignment (of the first 6 structures) was carried out with **COMPOSER** [Sutcliffe *et al*, 1987] and aligned with the other sequences using **MULTALIGN** [Barton, 1990] (with a structure dependent gap penalty of 20) using **DSSP** [Kabsch & Sander, 1983] to allocate secondary structure. This was displayed using **ALSCRIPT** [Barton, 1993].

1.5 Citrate Synthase

1.5.1 Metabolic Role

As discussed in section 1.3, the citric acid cycle is the final stage in the oxidation of foodstuffs to carbon dioxide, with the generation of NADH linked to the production of energy in the form of ATP. Secondly, the cycle also produces precursors for biosynthesis [Krebs *et al*, 1952]. After acetyl-CoA is formed from pyruvate, citrate synthase catalyses the condensation of oxaloacetate OAA and acetyl CoA to form citrate and CoA (below) involving the creation of a carbon-carbon bond and facilitating the entry of carbon into the citric acid cycle. It is therefore regarded as the first enzyme of the cycle, and thus plays an important regulatory role [Weitzman & Danson, 1976].

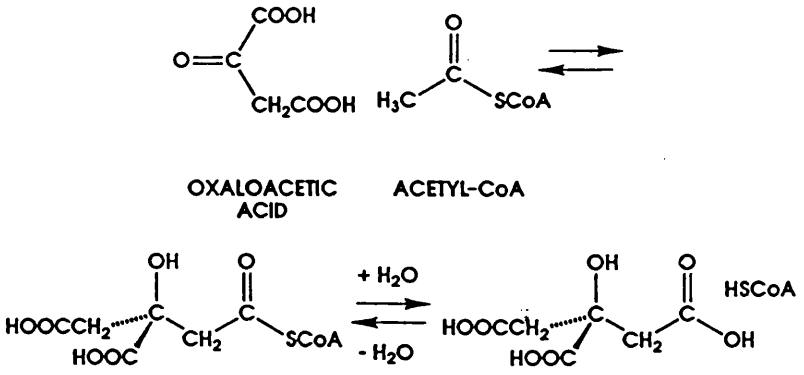


Figure 1.11: The reaction catalysed by citrate synthase [Remington, 1992b].

1.5.2 Structure

Initial analysis of the sizes of citrate synthase from various sources revealed that they can be divided into two groups; large and small [Weitzman & Dunmore, 1969]. The small form of the enzyme found in Eukarya, gram-positive Bacteria and Archaea is homo-dimeric in structure, however, gram-negative bacteria contain the large citrate synthase which is hexameric [Weitzman & Danson, 1976]. In all cases, the monomer size is very similar,

(M.W.= 40-50kDa), with both large and small enzymes containing 'catalytic dimers'.

As with TIM there is a very large sequence database and crystal structures determined for several organisms. Crystal structures are available for several complexes and crystal forms of the liganded (closed) and unliganded (open) forms of the pig heart [Remington *et al*, 1982] and chicken heart [Remington *et al* 1982, Liao *et al*, 1991] enzymes. There are also structures determined for the thermophile *Thermoplasma acidophilum* [Russell *et al*, 1994], the hyperthermophile *Pyrococcus furiosus* [Russell *et al*, 1997], and the psychrotolerant bacterium, DS23R [Russell *et al*, 1998].

Crystal structure analysis has shown that citrate synthase is almost entirely an α-helical protein, with only a small region of β-sheet. Each monomer comprises a large and small domain. As the structure of the *Sulfolobus solfataricus* enzyme is reported here, it is described in more detail in chapters six and seven, and compared with the previously determined structures (above).

1.5.3 Reaction Mechanism

The catalytic mechanism has been studied using site-directed mutagenesis [Hanford *et al* 1988, Kurz *et al*, 1992]. Studies on the pig enzyme have shown that it is only active as a dimer, due to the fact that active site residues reside on both monomers. Structures of the liganded form of pig citrate synthase showed that the substrates bind in a cleft between large and small domains, with large conformational changes taking place on substrate binding. The initial binding of oxaloacetate, initiates the movement of the small domain relative to the large, giving the closed form of the enzyme, and this coincides with a twenty-fold increase in binding constant for acetyl-CoA [Johansson & Peterson, 1974].

The reaction mechanism is thought to proceed via an acid/base mechanism [Remington, 1992a & b]. It does this by an initial enolisation (deprotonating the

methyl group of acetyl-CoA) followed by a condensation reaction in which the carbanion attacks the carbonyl group of oxaloacetate, producing citryl-CoA. The final stage involves a thioester hydrolysis resulting in the formation of products. The rate limiting step is the formation of the enol intermediate of acetyl-CoA. This mechanism is shown in figure 1.12. The catalytic residues involved in the reaction in the pig enzyme are thought to be histidine 274, (residing on the large domain) histidine 320, and aspartate 375 (from the small domain). Of the total 11 residues involved in substrate binding and catalytic action of the pig citrate synthase, 8 are conserved in the known Archaeal enzymes. In the initial concerted process, histidine 274 acts as an acid, protonating the carbonyl oxygen (on acetyl-CoA) while aspartate 375 acts as a base to deprotonate the methyl group. The neutral enol intermediate also reacts via an acid/base catalysis interaction (with histidine 274 acting now as the base and histidine 320 as the acid) with the formation of the carbon-carbon bond by attack on the carbonyl group of oxaloacetate. A proposed mechanism for the final hydrolysis stage of citryl-CoA involves attack of the thioester group by an activated water molecule, with concerted deprotonation followed by protonation of histidine 274. The initial base abstracted proton is then transferred to the CoA to give the final products [Remington, 1992a]. There is debate over the origin of the active base in this stage.

It is of interest to note that analysis of the V_{MAX} values for the pig, *T. acidophilum* and *P. furiosus* citrate synthases at their operating temperature (281± 23 for pigCS at 37°C, 100± 14 for *Ta*CS at 60°C, 226± 12 for *Pf*CS at 90°C (units being μmol product formed per minute per mg protein) show that the increase in stability may be at the expense of catalytic activity [Russell *et al*, 1997].

The different types of citrate synthase (large and small) also demonstrate alternative methods for regulation of energy production; the small citrate synthases have been shown to be regulated by isosteric inhibition of ATP, whereas the large citrate synthases are generally inhibited allosterically by NADH [Weitzman & Danson, 1976]. In contrast to this control via nucleotide inhibition, facultatively anaerobic organisms (where the cycle is operative

purely for biosynthesis) are inhibited by 2-oxoglutarate [Weitzman & Dunmore, 1968].

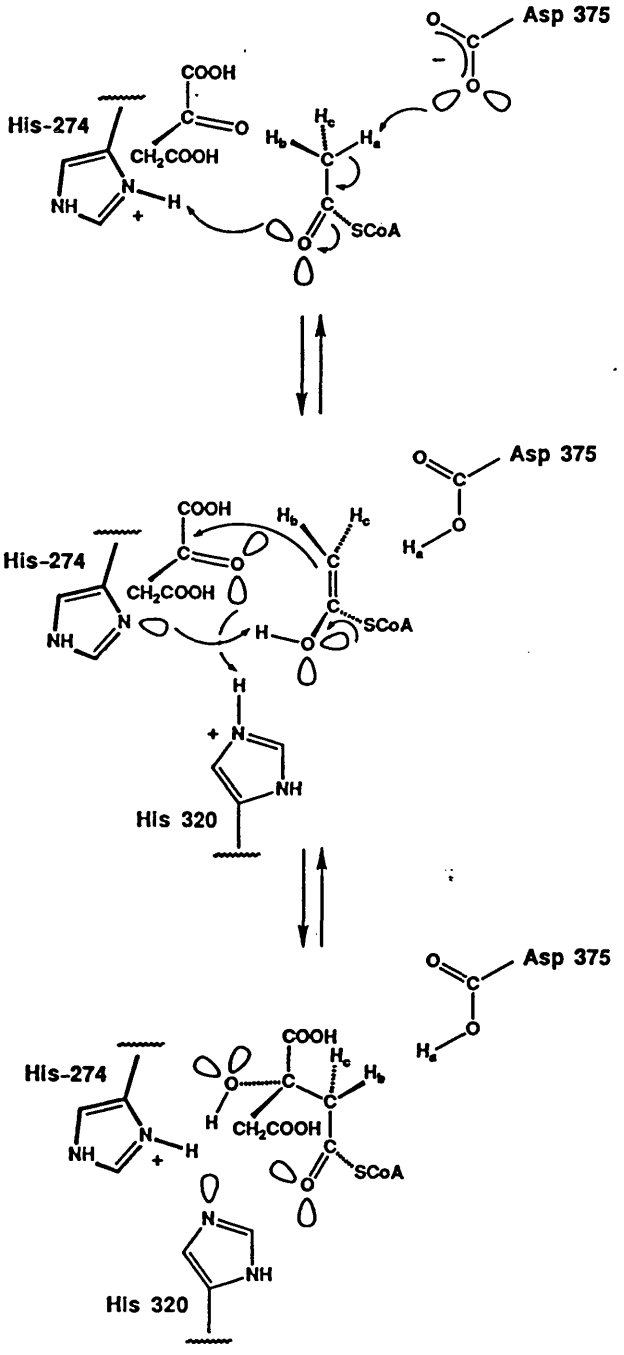


Figure 1.12a: Catalytic Mechanism of the condensation reaction catalysed by citrate synthase [Remington, 1992b].

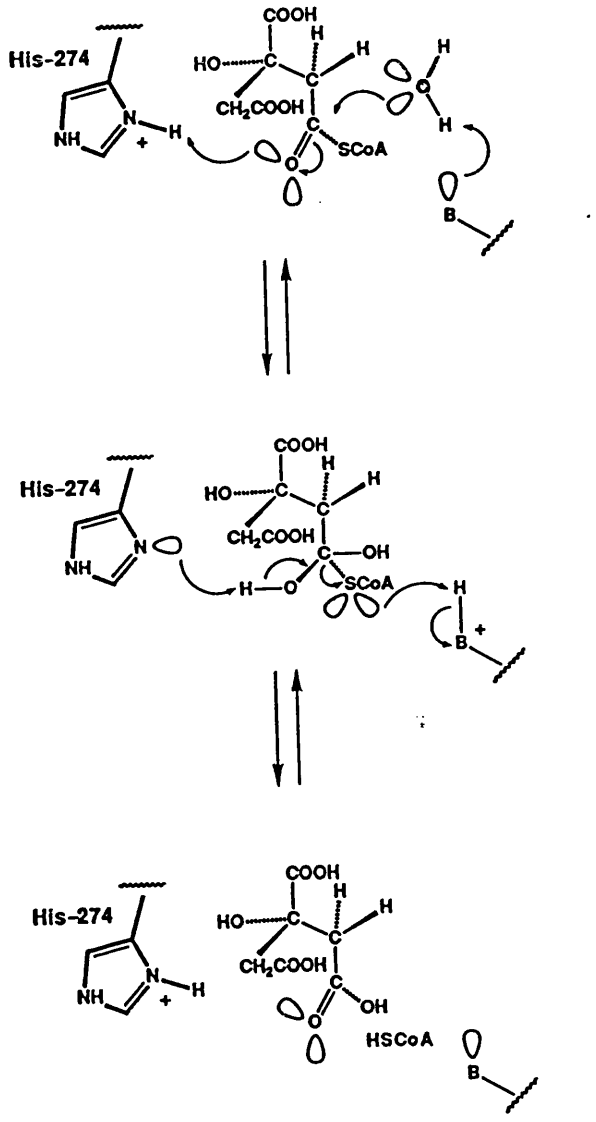


Figure 1.12b: Catalytic Mechanism of the hydrolysis reaction catalysed by citrate synthase [Remington, 1992b].

1.5.4 Citrate Synthase from *S. solfataricus* and Homology with known CSs

The gene for *S. solfataricus* citrate synthase was cloned, sequenced and overexpressed in *E. coli*, with the simple purification of a heat step followed by a dye-ligand affinity chromatography column [Connaris *et al*, 1998]. The amino-acid sequence comprises 378 residues, (Monomer MW. = 43 kDa) being 37 residues shorter than the pig enzyme, but more similar in length to the other Archaeal citrate synthases. The enzyme is also homo-dimeric, as with the other Eukaryal, Archaeal and gram positive Bacterial enzymes. In general, citrate synthase sequence identities between domains are between 20-30% (although the Bacterium DS23R has an identity of 41% with the Archaeal sequence *from P. furiousus*) and identities within domains tend to be in the range of 45-60%. Of the structural citrate synthases, the *T. acidophilum* sequence displays the highest sequence homology with that *S. solfataricus* sequence (with an identity of 57%) and as the former crystal structure has already been determined [Russell *et al*, 1994], the *S. solfataricus* structure was therefore expected to be very similar (again, a more thorough sequence and structural analysis is carried out for the structural citrate synthases in chapter seven, and a structure based sequence alignment of the structural citrate synthases can be found in figure 7.6).

1.6 Structural Features Involved in Protein Thermostability

1.6.1 Introduction

We have chosen to study protein thermostability by making comparisons from crystal structures of both mesophilic and thermophilic proteins. This work can also be carried out in conjunction with Site Directed Mutagenesis (SDM) experiments to remove supposed stabilising interactions from heat stable proteins, with the ultimate aim to engineer stability into mesophilic proteins. The main forces involved in protein folding have been reviewed [Dill, 1990]. There is only a marginal difference in free energy (ΔG_{FOLD}) between folded and unfolded states of any protein at a given temperature, T (below) ; typically 5-15

kcal/mol protein [Privalov, 1979], reflecting the balance between the need for protein stability and for enzyme turnover *in vivo*.

$$\Delta G_{FOLD}(T) = G_{NATIVE}(T) - G_{DENATURED}(T)$$

Stabilisation of the protein is achieved by many cumulative weak forces, including hydrogen bonds, hydrophobic, electrostatic, and Van der Waals interactions and the main opposing force is the conformational entropy of the protein chain. Altering only a few subtle interactions can therefore make a difference to the overall stability [Mathews, 1993]. Figure 1.13 [Dill *et al*, 1989] shows the temperature dependence of the hydrophobic effect and conformational entropy. The hydrophobic effect is thought to be the major driving force [Kauzman, 1959] for protein folding, and its contribution to ΔG_{FOLD} (or ΔF_T using the convention of Dill *et al* in figure 1.13) is shown to increase towards a maximum (most negative) at around 100°C. The conformational entropy is positive and has little temperature dependence at lower temperatures, but increases at higher temperatures. It is likely that heat denaturation is mainly driven by the increase in conformational entropy of the protein chain at higher temperatures, whereas cold denaturation is likely to be driven by a weakening of hydrophobic interactions at lower temperatures [Dill *et al*, 1989].

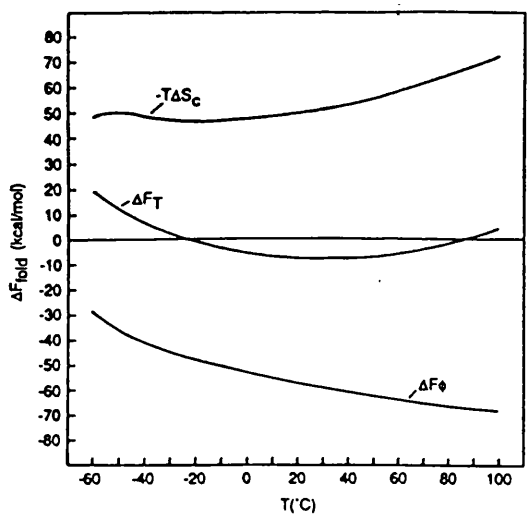


Figure 1.13: Calculated conformational entropy ($-T\Delta S_c$) and hydrophobic ($\Delta F\phi$) contributions to the free energy of folding (ΔF_T) as a function of temperature [Dill *et al*, 1989].

Since the initial studies were carried out on the crystal structures of ferredoxins [Perutz & Raidt, 1975] there has been a vast increase in the number of protein structures from thermophilic organisms in recent years. The observations made on elucidation of these thermophilic structures are of particular value if they can be compared to equivalent proteins from mesophilic systems. However, caution is required when interpreting observations as it is often assumed that thermal stability corresponds directly to the growth temperature of the organism, and this is not always the case [Vihinen, 1987]. The overall fold of a given protein obtained from different sources, tends to be very similar and it is therefore a combination of many subtle effects which may accumulate to give a protein tolerant to extremes of temperature. Several predominant trends have been observed in the structures of thermostable proteins, and their correlation to protein thermostability has been reviewed extensively [Russell & Taylor, 1995, Goldman, 1995, Vieille & Zeikus, 1996, Rees & Adams, 1995, Jaenicke *et al*, 1996]. Particular proteins have been shown to exhibit the various stabilising features to differing degrees.

Many of the observations are involved with improvement of packing of hydrophobic core or increase in the chain rigidity (particularly in moderate thermophiles). It is obvious however that a balance must be obtained between the increased compactness or rigidity necessary for increased thermostability, with the need for a certain amount of flexibility which must be retained in order for a protein to perform its functional or structural role *in vivo*. In hyperthermophilic proteins, electrostatic interactions are emerging as the most common mechanism by which increased stability is achieved [Vogt *et al*, 1997, Spassov *et al*, 1995].

Differences at the amino acid level obviously dictate differences in tertiary structure and the discussion below concentrates on some of the observations made from both sequence, crystal structure and SDM studies, and the implications to protein stability with respect to the forces involved.

1.6.2 Amino-Acid preferences in thermophilic proteins

Many of the initial studies (particularly in the absence of many 3-D structures) concentrated on the preferred amino-acids in thermophilic proteins. Menendez-Arias and Argos proposed hot-to-cold exchanges observed in six protein families with results suggesting main stabilising features being due to decreased flexibility and increased hydrophobicity in areas of secondary structure, particularly helices and subunit interfaces [Menendez-Arias and Argos, 1989]. Amongst the most noted exchanges were that of lysine to arginine and several other substitutions leading to 'alanine rich' helices which increase rigidity maximising hydrophobic contacts (e.g. gly→ala, ser→ala, lys→ala, thr→ala). The lysine to arginine exchanges were previously quantified as an increase in R/(R+K) ratio with increasing growth temperature of organism [Merkler *et al*, 1981]. Mutation of lysine to arginine in several proteins has been shown to increase stability, and it is proposed that this is due to the increased hydrogen bonding potential of the guanidinium group [Mrabet *et al*, 1992]. Zuber *et al* also suggested an exchange of polar amino acids in mesophilic enzymes for hydrophobic and charged residues in thermophiles [Zuber et al, 1988]. These preferences are likely to be site specific [Russell & Taylor, 1995].

A general reduction in the number of 'thermolabile' residues with increase in temperature has been observed in some proteins [Menendez-Arias & Argos, 1989, Russell & Taylor, 1995], mainly due to the possibility of oxidation of cysteine and methionine and deamidation of asparagine and glutamine at high temperatures. Studies on deamidation of asparagine and glutamine residues have suggested some of the other stereo-chemical features involved, [Kossiakoff, 1988, Daniel *et al*, 1996] and the importance of the neighbouring side chains which can alter the likelihood of formation of the cyclic imide intermediate neccesary for deamidation to occur.

Other specific amino-acid preferences are discussed in the following sections with regard to specific interactions.

1.6.3 Stability of alpha helices

The stability of alpha helices has been much studied in terms of the propensities of particular amino-acids for different intra-helical positions [Menendez-Arias & Argos,1989, O'Neil & Degrado, 1990 , Serrano *et al* 1992, Horovitz *et al* 1992] and preferred helix 'capping' residues, particularly where stabilising charge-dipole interactions are present [Nicholson *et al*, 1991, Richardson and Richardson, 1988, Forood *et al* 1993, Fersht & Serrano, 1993]. There is also evidence to suggest that helices are stabilised by i(i+3) and i(i+4) glu-lys salt bridges [Scholz *et al*, 1993, Lyu *et al*, 1992]

Several other proteins have shown high extent of helix stabilisation, (e.g. increase of alanines in helices in thermophile [Kelly *et al*, 1993]) however, a recent study on the helices of thermophilic proteins has demonstrated that only one trend was conserved across different protein families where crystal structures were currently available; the absence of beta-branched residues at an intra-helical position due to the unfavourable torsion angles which they introduce [Facchiano *et al*, 1998].

1.6.4 Conformational Entropy of the Unfolded State

Many of the methods by which a protein may achieve stabilisation may seek to lower the conformational entropy of the unfolded protein. Shortening of the protein chain (which is most likely to occur in loop regions, due to a need to conserve secondary structure) should also serve to decrease the conformational entropy by restricting the number of conformational states of the unfolded protein. The conformational entropy of a protein may also be reduced by rigidifying the chain; the introduction of proline residues into loop regions and beta turns via site-directed mutagenesis [Mathews *et al* 1987, Hardy *et al*, 1993, Watanabe *et al* 1991, 1994, 1997, Bogin *et al*, 1998] has been shown to have stabilising effects. Watanabe *et al* have also shown this effect to be cumulative [Watanabe *et al* 1994] and confirmed deductions with analysis of the crystal structure [Watanabe *et al* 1997]. Prolines at an internal helical position however, break at least two hydrogen-bonds and thus should

destabilise a helix [Richardson & Richardson, 1988)]. An increase in the number of proline residues has been observed in loop regions of other thermophilic structures [Delboni *et al*, 1995, Russell *et al*, 1998]. Conversely, glycine should have a destabilising effect due to the increased flexibility introduced [Horovitz *et al*, 1992, Mathews *et al*, 1987]. Imanka *et al* demonstrated an increased thermostability in *B. stearothermophilus* neutral protease via a single glycine to alanine replacement [Imanaka *et al*, 1986]. However, this situation is complicated by the fact that a glycine residue may be present in a position which allows the mainchain to adopt a conformation which will result in better packing or other more favourable interactions in the folded protein [Korndorfer *et al*, 1995, Ishikawa *et al*, 1993]. In addition it has been indicated that glycine is favourable at the C-termini position of helices where it satisfies the H-bonding potential of the mainchain N-H groups by allowing exposure to solvent [Richardson & Richardson, 1988].

1.6.5 Packing Efficiency

The overall compactness of a protein can be reflected by a low surface area to volume ratio, [Chan *et al*, 1995] increased packing density [Britton *et al*, 1995] with fewer solvent exposed residues and internal cavities (or total cavity volume) in addition to shorter loop regions when compared with the mesophilic protein. This should result in improved hydrophobic packing in the core of the protein (see 1.6.6) and better Van der Waals interactions [Privalov & Gill, 1989].

Loop regions have a tendency towards higher mobility than regions of secondary structure and molecular dynamics simulations have proposed that turn and loop regions are likely to be initiating sites for protein unfolding [Dagget & Levitt, 1993] suggesting a need for either shortening or additional stabilisation in thermophilic proteins. The shortening of loop regions between areas of secondary structure has been observed in thermophilic citrate synthases [Russell *et al*, 1994 & 1997].

Studies have been carried out on high resolution protein structures to investigate the occurrence and role of internal cavities in proteins [Hubbard *et al*, 1994]. Cavities have been shown to be almost always present in proteins over ~100 residues in size, and in general, overall cavity volume has been shown to increase with protein size (but only occupying up to ~2% of the total protein volume). The loss of Van der Waals interactions due to the presence of cavities has been investigated in Site-directed mutagenesis experiments with some results confirming the destabilising effects of cavities [Kellis *et al* 1988, 1989, Eriksson *et al* 1992] and Eriksson *et al* suggest the value being around 24-33 calmol^{-1}Å^{-3}. Some experiments have however failed to prove the significance of these features [Eijsink *et al* 1992]. Cavities can be either empty, showing a typically hydrophobic character, or contain solvent and exhibit more polar surfaces, which facilitate hydrogen-bonding with buried waters. It has therefore been suggested that they can act to stabilise these buried polar residues and thus helping to accommodate such groups in the protein core and stabilise particularly larger cavities [Hubbard *et al*, 1994]. Some enzymes show little difference in the number of internal cavities between mesophile and thermopile [Wallon *et al*, 1997] but other proteins have shown a certain trend towards either fewer cavities or reduced cavity volume in the thermophile [Russell *et al*, 1994, Delboni *et al*, 1995].

1.6.6 Hydrophobicity

Burley and Petsko suggested the importance of aromatic-aromatic interactions in protein stability [Burley & Petsko, 1985], with pairwise interactions contributing -1 to -2 kcalmol^{-1}, and the example of a mutant λ repressor protein, introducing a tyrosine residue which increased the thermostability by 6°C compared with wild type enzyme due to the formation of an aromatic-aromatic network. An increase in the extent of aromatic interactions has also been observed in the crystal structures of both *T. thermophilus* Ribonuclease H [Ishikawa *et al*, 1993] and thermitase from *T. vulgaris* [Teplyakov *et al*, 1990]. Kellis *et al* demonstrated that the truncation of a single methylene group in barnase resulted in a destabilisation of 1.1 kcalmol^{-1} [Kellis *et al*, 1988] and an increase in 'aliphatic index' in thermophilic proteins has also been observed

[Ikai, 1980]. Increased hydrophobic interactions at subunit interfaces of thermophiles are thought to be of significance in several crystal structures (see 1.6.10). An increase in isoleucine content which manifests itself in the form of clusters in the protein core, was also noticed in *P. furiosus* GLUDH, and it has been suggested that isoleucine may be favoured over leucine due to the ability to exist in all three $\chi 1$ angles compared with predominantly two for leucine, allowing more efficient packing in the protein core [Yip et al, 1995].

1.6.7 Disulfide bridges

The covalent crosslinking between two cysteine sulphydryl groups should stabilise a protein via a reduction in the entropy of the unfolded state, thus entropy favouring the folded protein (see 1.6.4). The increased stability of several proteins has been demonstrated via engineering in cysteine residues to form disulfide bridges, [Matsumura, 1989, Van den Burg *et al*, 1988] however an engineered disulfide bridge can introduce unfavourable strain energy and in addition, the lower protein-water hydrogen bonding of a 'bridged' protein can reduce the entropy gain due to exclusion of water on protein folding [Goldman, 1995].

1.6.8 Hydrogen bonding

The strength of a hydrogen-bond has been demonstrated [Shirley *et al*, 1992] with this strength shown to increase if either the donor or acceptor is charged [Fersht, 1985]. It has been suggested that hydrogen-bonding may not play such an important role in overall protein thermostability as there is probably no significant difference in the total energies of protein-water hydrogen bonds of the unfolded state compared with protein-protein interactions in the folded protein [Dill, 1990]. It is accepted however that these bonds are important in forcing directional constraints on the structure. More recently however, a study by Vogt *et al* have suggested the widespread importance of hydrogen-bonding in thermophilic protein structures, with many of these proteins displaying increased fractional polar surfaces than their mesophilic counterparts [Vogt *et*

al, 1997]. In addition, a single alanine to serine mutation in *B. stearothermophilus* neutral proteinase increased stability through the introduction of an internal hydrogen-bond [Eijsink *et al*, 1992] and several researchers have proposed that hydrogen-bonding may play a role in the thermostability of other structures [Paupit *et al*, 1988, Hennig *et al* 1997, Tanner *et al*, 1996].

1.6.9 Electrostatic interactions

Initial structural studies by Perutz and Raidt [Perutz & Raidt, 1975] led to the suggestion that surface ion pairs could confer thermostability in proteins. It has been suggested that the energy of a single ion pair contributes very little to the stability of a protein at mesophilic temperatures due to the fact that the entropic cost of desolvating and immobilising a charged residue is likely to be similar to the favourable electrotstatic energy gained from the charge-charge interaction [Mathews, 1993, Honig & Nicholls, 1995] (although this situation may differ slightly between protein folding and oligomer association as described in section 1.6.10). At hyperthermophilic temperatures however, stabilisation due to ionic interactions should increase due to the lower dielectric constant of water, and thus the strength of a surface ion pair is effectively greater at high temperatures [Elcock, 1998]. An ionic network should also be more favourable than a pairwise interaction, as a similar coulombic energy can be gained for a lower entropic cost [Horovitz *et al*, 1990].

This theory has been confirmed by the many structural observations revealing increased ionic interactions (particularly networks) in many hyperthermophilic proteins compared with their mesophilic counterparts, [Russell *et al*, 1997, Day *et al*, 1992, Hennig *et al*, 1995, Davies *et al*, 1993, Kelly *et al*, 1993, Korndorfer *et al*, 1995, Yip *et al* 1995, Lim *et al*, 1997, Hatanaka *et al*, 1997]. Electrostatic interactions are therefore emerging as one of the most singular methods of stabilisation, particularly in proteins from organisms growing at the upper end of the temperature scale.

There are now a number of studies which stress the importance of ionic interactions (as with hydrophobic interactions) at protein interface regions (see 1.6.10). Specifically, glutamate deydrogenase from *P. furiosus* [Yip *et al*, 1995, Rice *et al* 1996] and malate dehydrogenase from *T. flavus* [Kelly *et al*, 1993] have increased ionic interactions (relative to mesophile) at both inter-subunit and inter-domain interfaces (with the former displaying extensive ionic networks). Another site where ionic interactions are likely to play an important role are at the termini, where they are likely to prevent fraying [Day *et al*, 1992, Hennig *et al*, 1995, Starich *et al*, 1996, Russell *et al*, 1997]. Disruption of an intra-subunit ionic network involved with termini interactions in *Tm*GAPDH, resulted in a considerable reduction in thermostability [Pappenberger *et al*, 1997].

In some cases intra-molecular ionic interactions are also likely to be of importance; IMPDH from *T. thermophilus* displays an increase in intra-subunit ion pairs compared with the mesophile [Wallon *et al*, 1997] and Fe-superoxide dismutase from *A. pyrophilus* shows an increase in intra-subunit surface ionic interactions [Lim *et al*, 1997] in addition to ionic networks at the dimer interface.

The argument that ionic interactions are longer range, in contrast to the shorter range hydrophobic interactions (which have a $1/r^6$ dependence) has also been used to explain the more resilient strength they may give to a protein [Aguilar *et al*, 1997]. In the latter case (as in several proteins) ion pairs present at the surface of a protein (particularly stabilising flexible regions of surface structure) form a 'net' like structure, over the protein surface, which may prevent solvent penetrating the hydrophobic core [Walker *et al*, 1980].

1.6.10 Oligomerisation and importance of interface interactions

Comparisons of a number of mesophilic and thermophilic oligomeric protein structures have illustrated the importance of interface interactions in a role for protein stabilisation and are noticeably one of the areas of the structure which show the greatest differences. The thermophiles tend to show an increase in either hydrophobic interactions (through better packing or increased

hydrophobic surface buried) [Walker *et al*, 1980, Russell *et al*, 1994, Delboni *et al*, 1995, Hennig *et al*, 1997, Wallon *et al*, 1997, Lim *et al*, 1997, Knapp *et al*, 1997, Komdorfer *et al*, 1995] or in electrostatic and or hydrogen bonding at the interface [Yip *et al*, 1995, Tanner *et al*, 1996, Lim *et al*, 1997, Kelly *et al* 1993b, Russell *et al*, 1997]. The consequences of the differing solvation of polar and charged residues before and after both folding and subunit-subunit association, may also suggest why hydrophilic bridges are particularly favourable at subunit interfaces [Xu *et al*, 1997].

As the catalytic activity of an oligomeric enzyme generally depends on the integrity of the oligomer, these regions of the structure are likely to be a key in stabilisation, and it has been shown that the dimer to monomer transition is the first step in the denaturation process for pig citrate synthase [McEvily & Harrison, 1986]. Several thermophilic proteins may also achieve thermostability by existing in higher oligomeric states than their mesophilic counterparts. Increased thermostability may occur via a reduction in exposed surface area with increased burial of hydrophobic sidechains. The crystal structure of *P. furiosus* ornithine carbamoyltransferase [Villeret *et al*, 1998] has shown that the enzyme is a dodecamer, in comparison to the trimers of mesophilic Bacteria. In the latter case, the interface regions between catalytic trimers exhibited a high degree of hydrophobicity, with several buried isoleucine and tyrosine residues. TIM from both *P. woesei* and *M. fervidus* may also achieve stabilisation in a similar way, being tetrameric in comparison to the dimeric Eukaryal and Bacterial enzymes [Kohlhoff *et al*, 1996]. *T. maritima* PRAI [Hennig *et al*, 1997] and *S.solfataricus* β-glycosidase [Aguilar *et al*, 1998] also display that a move towards a higher state of oligomerisation is a recurrent feature in thermostable proteins.

1.6.11 Limitations of Structure Based Studies

It is clear from the above discussion, that there is no single preferred mode of protein stabilisation to extremes of temperature. In addition, the subtlety of interactions involved are highlighted by examples of structures in which there are few obvious features which could explain the large differences in

thermostability. One example is that of CheY from *T. maritima* [Usher *et al*, 1998] (comprising only 130 residues) which shows no difference in hydrogen-bonds, compactness, ion pairs or exposed hydrophobic surface, with only slight differences being the shortening compared with the *E. coli* enzyme and two additional proline residues. Some of these less obvious cases may be clarified with site-directed mutagenesis experiments, but these are of course not straightforward. It should also be noted that by analysis of crystal structures we are only studying interactions possibly stabilising the folded protein, and not differences in folding processes themselves. We also have limited information concerning these proteins in the dynamic environment of their normal operating temperatures. Lastly, although proteins from these organisms are inherently more stable than the equivalent from a mesophilic source, there may be other factors such as the high concentrations of intracellular metabolites (such as inositol) which play additional roles in stabilising proteins in thermophiles, [Scholz *et al*, 1992] and aspects such as these are beyond the experimental scope of this study.

1.7 Experimental Aims

The aims of the project were firstly to produce the recombinant *P. woesei* TIM which had been overexpressed in *E. coli* [Bell *et al*, 1998] and set up the purification process in house such that protein of a sufficient level of homogeneity could be obtained in order to carry out crystallisation studies. It was then hoped to obtain X-ray diffraction quality crystals of the TIM and subsequently attempt to determine the crystal structure, initially with the Molecular Replacement method using the structural data of the mesophilic and thermophilic TIMs. Difficulties with the Molecular Replacement problem, and the following search to acquire good heavy atom derivatives, led to the second project on *S. solfataricus* citrate synthase being initiated. The new aims therefore became to collect X-ray data, and determine the structure of this enzyme to allow an extension of the structural study on Archaeal citrate synthases which has been previously carried out within the group. This would allow the collation of all structural data, in order to carry out a complete

analysis of the citrate synthases from five organisms growing over a vast temperature range.

CHAPTER 2

P. woesei TIM; Purification and Crystallisation

2.1 Introduction

Prior to my involvement in the project, a system had been set up to express the *Pw*TIM in *E. coli* and a purification protocol established. This work was carried out by Michael Kohlhoff at the University of Essen, Germany. On receiving the clone, the protein purification procedure had than to be set up 'in house' to enable enough protein to be produced for crystallisation trials. This protocol was adapted and developed according to practical limitations such as the availability and performance of chromatography columns, and the process described below represents the final established method.

2.2 Overexpression of *Pw*TIM in *E. coli*

E. coli strain DH5α cells expressing the vector pJF118EH [Furste *et al*, 1986] were used for overexpression of the *P. woesei* TIM [Bell *et al*, 1998]. The culture medium used contained 10g yeast extract, 10g peptone, 5g NaCl, pH 7.3 for a 1 litre culture. After autoclaving, 50 μg/ml ampicillin was added to the medium. 1-2μl cells (glycerol stock stored at -20°C) were added to a 10ml broth and these cells grown up for 5-6 hours under aeration at 37°C. 1ml of this culture was then added to the 1litre broth (in a two litre conical flask) and the culture was grown to A_{578}=1.0 (again at 37°C). Induction was then carried out by addition of 200mg/litre IPTG, and cultivation continued over a further 6 hours. The medium was then centrifuged (20 minutes @ 5 000 G), the supernatant decanted off, and the wet cell pellet (approximately 5-6g of wet cells per litre) resuspended to a concentration of 0.2g/ml in 50 mM TRIS/HCl pH 8.5 containing 25mM NaCl, 2mM EDTA and 1 mM PMSF.

2.3 Purification of *Pw*TIM

The cells were lysed by sonication and then centrifuged (30 minutes @ 10,000G) and the supernatant collected. This crude cell extract was subjected to a heat step (15 minutes @ 85°C), in order to denature *E. coli* host proteins. The sample was then centrifuged (30 minutes @ 10,000G) and again the supernatant decanted off, and collected. This heat treated extract was then passed through a 0.2μm filter (using a 20ml syringe) before loading onto an anion exchange column (Pharmacia Sephacryl 300, 400 ml column) by means of a 150ml 'superloop', after first equilibrating the column with 50 mM TRIS/HCl pH 8.5 containing 25mM NaCl and 2mM EDTA (buffer A). Several bed volumes of buffer A were run through the column to wash off unbound material, (no TIM activity was present in the unbound material) and the TIM eluted off with a sodium chloride gradient (0M-0.1M in 50mls and 0.1M-0.3M 50-450mls, flow rate/min). 10ml fractions were collected on starting the salt gradient, and most of the TIM activity was eluted off at an NaCl concentration of 0.15M (figure 2.1, fractions 11-18). These fractions (comprising 80mls) displaying TIM activity were pooled and concentrated in an Amicon centreprep-30® concentrator to a total volume of 1ml for the gel filtration step. Remaining bound protein was removed from the column by washing with a 2M NaCl solution. The gel filtration column (Pharmacia Superdex 200) was equilibrated in buffer A (above) and the 1ml sample loaded onto the column and run at a flow rate of 1ml/min. 40 mls of buffer A were run through the column (approximately equal to the void volume) before collecting 1ml fractions, and the bulk of the activity was eluted off after approximately 75mls (figure 2.2, fractions 32-42). The pure TIM fractions were pooled (total 11ml) and centrifuged in an Amicon centricon-30® to a final concentration suitable for crystallisation trials (typically 10 mg/ml). The purity of the *Pw*TIM was monitered throughout by means of measuring specific activities using the assay (section 2.4) allowing construction of a purification table as seen in figure 2.5 and SDS-PAGE gels (section 2.5).

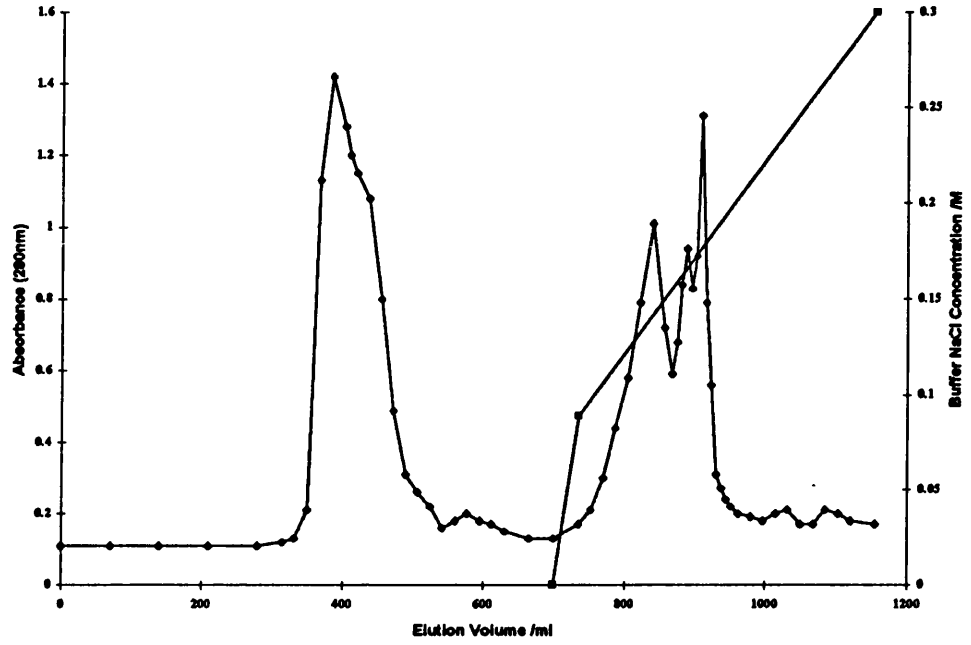


Figure 2.1: Anion Exchange chromatography (with the salt gradient also shown).

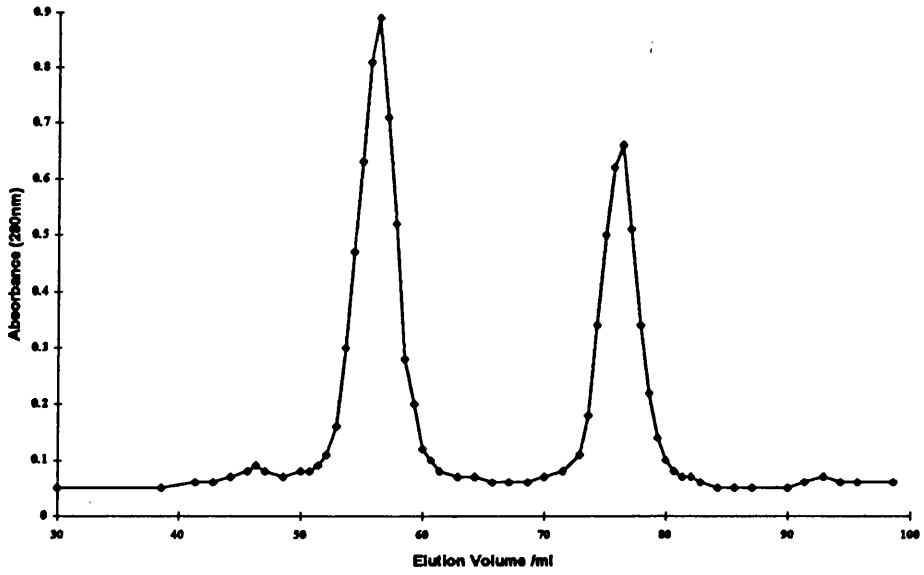


Figure 2.2: Gel Filtration chromatography.

2.4 Spectrophotometric Assay for Measuring *Pw*TIM Activity

In order to measure the activity of the *Pw*TIM an enzyme coupled assay was used, with glyceraldehyde-3-phosphate dehydrogenase (GAPDH) as the auxiliary enzyme. The increase in NADH concentration was measured at 340nm thus effectively monitoring the conversion of dihydroxyacetone phosphate (DHAP) to D-glyceraldehyde-3-phosphate (D-GAP). The assay was that of Kohlhoff *et al* [Kohlhoff *et al*, 1996], with the exception that it was performed at 55°C (not 70°C) as the GAPDH was obtained from rabbit muscle (SIGMA) and therefore not stable for sufficient time above this temperature. The assay mixture contained 100mM TRIS/HCl (pH 7.4 @ 55°C), 5mM sodium arsenate, 10mM NAD^+, 4mM dihydroxyacetone phosphate and 20U glyceraldehyde-3-phosphate dehydrogenase (all from SIGMA) at a total volume of 1ml. The system is described by figure 2.3 which shows that addition of Arsenate results in the irreversible nature of the reaction catalysed by GAPDH allowing the desired rate to be measured.

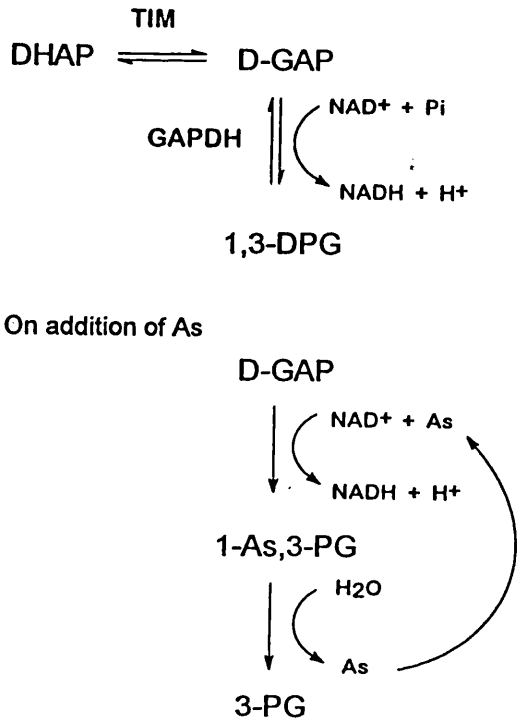


Figure 2.3: Enzyme assay used to monitor purification procedure (As = arsenate).

Although this assay system had already been established, it was important to check firstly that the coupling enzyme (GAPDH) was present in high enough quantities such that it was not rate limiting. This was done simply by adding twice the amount of this enzyme and ensuring that no increase in measured rate occurred. In addition, as the assay was performed at 55°C the stability of this rabbit enzyme over the time course of the experiment was found to be sufficient, being stable at the above temperature for 2.5 minutes (and initial rates typically measured over 30 second periods). Finally, adding either half or double the amount of *Pw*TIM to the assay was shown to have a similar effect on the measured rate, and therefore it was assumed that initial (linear) rates were in fact being measured (with no rate observed on elimination of TIM from the assay).

It should be noted that when calculating activity values for *Pw*TIM, these should really be corrected for the temperature dependent hydration of both the substrates which alter the amounts of the active form of each substrate present in solution (and hence affecting data calculated for the interconversion of GAP-DHAP) [Trentham *et al* 1969, Reynolds *et al* 1971]. The correction was not applied in this case as extremely accurate kinetic data were not required for monitoring the purification process.

Basic characterisation has been carried out on the recombinant *Pw*TIM (by Helen Lawrence) and thermal stability data were in accordance with those reported by Kohlhoff *et al* [Kohlhoff *et al*, 1996] (the latter being collected for TIM purified from *P. woesei*).

2.5 Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis (SDS-PAGE)

Protein homogeneity was monitored at different stages of the purification process using the method described by Laemmli [Laemmli, 1970] with a 5% (w/v) stacking gel and 12.5% (w/v) resolving gel. Biorad® standard molecular weight markers were used; 14.4 kDa, 21.5 kDa, 31 kDa, 45 kDa, 66.2 kDa, 97.4 kDa.

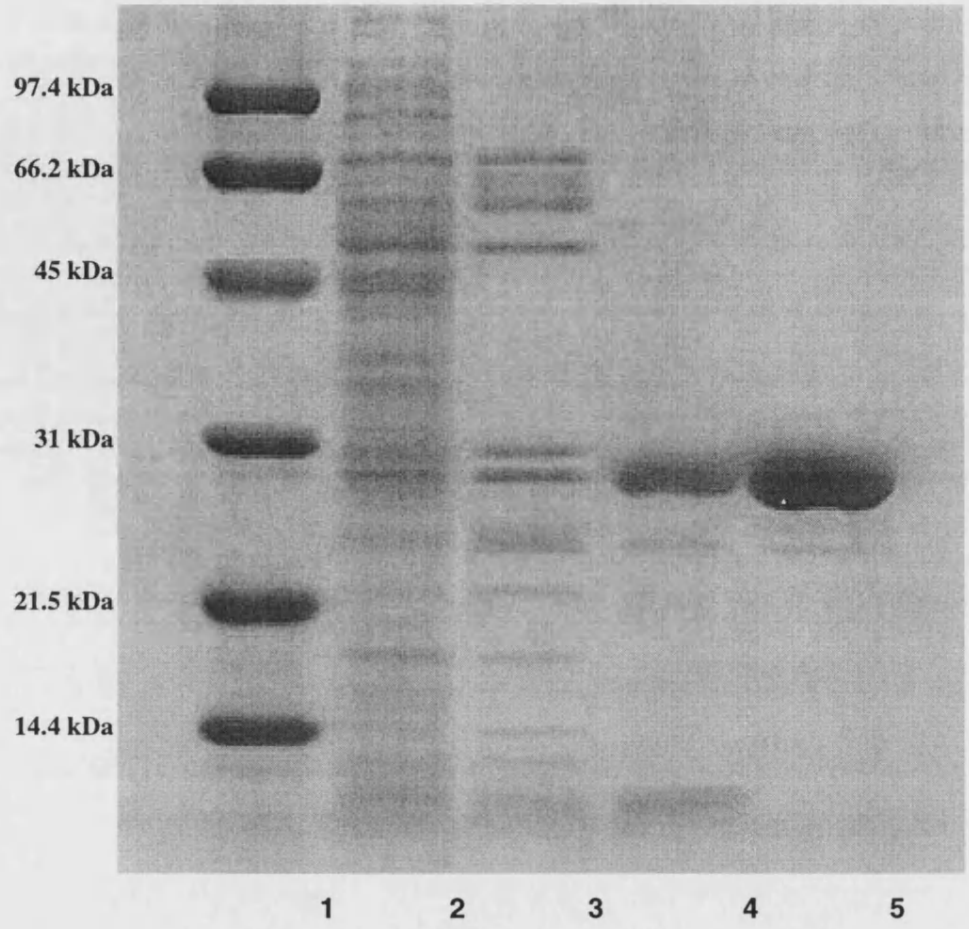


Figure 2.4: SDS-PAGE of samples from each stage of the purification process.

Lane 1: Molecular Weight Markers.

Lane 2: Crude Cell Extract.

Lane 3: Heat Treated Extract.

Lane 4: Pooled Anion Exchange

Lane 5: Pooled Gel Filtration.

2.6 Determination of Protein Concentration by Bradford Assay

The absorbence of a 1ml sample containing 100µl protein solution and 900 µl Bradford reagent (0.01% (w/v) coomassie Blue G-250, 5% (v/v) ethanol and 8.5% (v/v) H_3PO_4) was measured at 595nm after incubation at room temperature for 15 minutes [Bradford, 1976]. Protein concentration was

calculated by comparison with a bovine serum albumin (BSA) standard curve, constructed for a concentration range of 0-200 μg/ml.

Purification step	Total Protein /mg	Total Activity /U	Specific Activity /(U/mg)	Purification /Fold	Yield /%
Crude Extract	487	13516	28	—	100
Heat Step	54.4	9145	168	6.0	68
Anion Exchange	10.3	7019	681	24.3	52
Gel Filtration	5.6	4595	821	29.3	34

Figure 2.5: Purification of *P. woesei* TIM expressed in *E. coli*, (1 litre cell culture; 5.8g wet cells).

2.7 Crystallisation of Recombinant *Pw*TIM

All crystallisation trials were set up using the hanging drop vapour diffusion method [Jancarik & Kim, 1991]. In order to screen for conditions suitable for growth of crystals, initial trials were set up using the sparse matrix Hampton Research Crystal Screen™ I and II reagent kits, comprising a total of 98 different conditions known to have yielded protein crystals. These trials were initially performed on the native enzyme at a protein concentration of 10 mg/ml, in a stock solution of 50 mM TRIS/HCl pH 8.5 containing 25mM NaCl and 2mM EDTA. Initially, two of the 98 Crystal Screen conditions gave small needle like crystals; after 48 hours in condition 42 of Crystal Screen I (0.05M potassium phosphate 20% PEG 8K), and also after 2 weeks in condition 36 of Crystal Screen I (0.1M TRIS/HCl pH 8.5, PEG 8K). Additional matrices of varying pH and PEG concentrations were set up around these crystal yielding conditions, but although some larger needles and small rod like crystals were obtained, they proved unstable and were not of a quality suitable for X-ray diffraction.

Following these trials, attempts were made to co-crystallise *P. woesei* TIM with a substrate analogue, 2-phosphoglycolate (2-PG) present in mM concentrations. Condition 37 of Crystal Screen I (0.1M sodium acetate pH 4.6, 8% PEG 4K) yielded small rod like crystals after a period of 12 hours. Improved crystals, still of a rod like habit were obtained at slightly lower PEG concentrations (5-7%), grown over a period of about 5 days. Attempts were made to optimise this condition by making a number of alterations; varying the percentage of precipitant (polyethylene glycol (PEG)) and also using PEG of different molecular weight (2K, 6K, 8K and 20K), pH of the mother liquor, changing protein concentration (between 5 and 20 mg/ml) and altering the amount of 2-PG between 10mM-100mM. A further additive screen was used comprising a variety of organic reagents (PEG, MPD, alcohol, glycerol) and cations (metal chlorides e.g. $CaCl_2$) in an attempt to alter the rate of diffusion or create additional interactions affecting crystallisation; none of these were successful. The hanging drop size was also varied (4μl,6μl,8μl) but seemed to have little effect on crystal size. In addition, the effect of temperature on crystallisation was investigated by carrying out trials at 48°C and 4°C, in addition to those at room temperature. All conditions at 48°C in which crystals were obtained, displayed excessive nucleation (many small needles), however rod like crystals were grown at 4°C, with slightly higher PEG concentrations (8-9%), grown over a period of 1 week. The best crystals were obtained at room temperature with 20mM 2-PG, 0.1M NaAc pH 4.2 , 5%PEG 4K grown over a period of 1 week, with a protein concentration of 10 mg/ml and the TIM/2-PG mixture incubated for 1 minute at 50°C. These crystals were found to be of an orthorhombic space group ($P2_12_12$) and the best diffraction achieved was to a resolution of 2.9Å. It was also noticed that *Pw*TIM would crystallise in a hexagonal rod habit under the same conditions (cohabiting the same drop) but these crystals were later shown to be disordered. In addition to condition 37 of Crystal Screen I, five conditions in Crystal Screen II yielded small needle-like crystals (No. 22; 0.1M MES pH 6.5, 12% PEG20K, No. 30; 0.1M HEPES pH 7.5, 10% PEG 6K, 5% MPD, No. 37; 0.1M HEPES pH 7.5, 10% PEG 8K, 8% ethylene glycol, No. 38; 0.1M HEPES pH 7.5, 20 % PEG 10K and No. 45; 0.1M TRIS/HCl pH 8.5, 0.01M Nickel Chloride Hexahydrate, 20% PEG monomethyl

ether 2K). Initial matrices set up to optimise these conditions provided no increase in crystal size.

In a further attempt to achieve better diffracting crystals, several other substrate/transition state analogues were used, for co-crystallisation; 2-carboxyethyl phosphonic acid (2-CP), and glycerol-3-phosphate. No crystals were obtained using glycerol-3-phosphate, however, TIM complexed with 2-carboxyethylphosphonic acid, yielded several crystals more suitable for X-ray diffraction studies. These crystals grew in a similar condition, found to be optimum for TIM/2-phosphoglycolate (20mM 2-CP (incubated with TIM for 1minute at 55°C), 0.1M NaAc pH4.0, 7%PEG 4K, grown at room temperature over 5 days). These crystals diffracted to a resolution of 2.6Å, and were found to belong to the monoclinic space group ($P2_1$). Further attempts to optimise the crystallisation condition varying the parameters above (as was carried out for 2-PG) failed to improve X-ray diffraction quality.

2.8 Discussion

The purification process described allowed 5-6mg of pure TIM to be produced per litre culture of cells which was deemed sufficient yield to create a supply for crystallisation studies. Specifically, expression of the thermostable *Pw*TIM in a mesophilic host allowed the use of a heat denaturation step which greatly simplified the procedure. The SDS-PAGE also demonstrates protein homogeneity throughout the purification process with the overloaded sample after the gel filtration run suggesting the TIM to be sufficiently pure with the presence of only a couple of faint minor bands. Further to this, the specific activity measured for the pure *Pw*TIM was 821U/mg at 55°C (Figure 2.3). If the activity should approximately double for an increase in temperature of 10°C [Danson *et al*, 1996] this figure is slightly lower, although comparable with that reported by Kohlhoff [Kohlhoff *et al*, 1996] (3900U/mg) given that the assay in that case was carried out at 70°C (closer to the optimum temperature of activity for *Pw*TIM). It is possible that experimental error included in measurement of enzyme rates (such as actual cuvette temperatures) and protein concentration assays could account for the difference. A further high resolution anion exchange column (such as Pharmacia Mono-Q) could probably be used as a

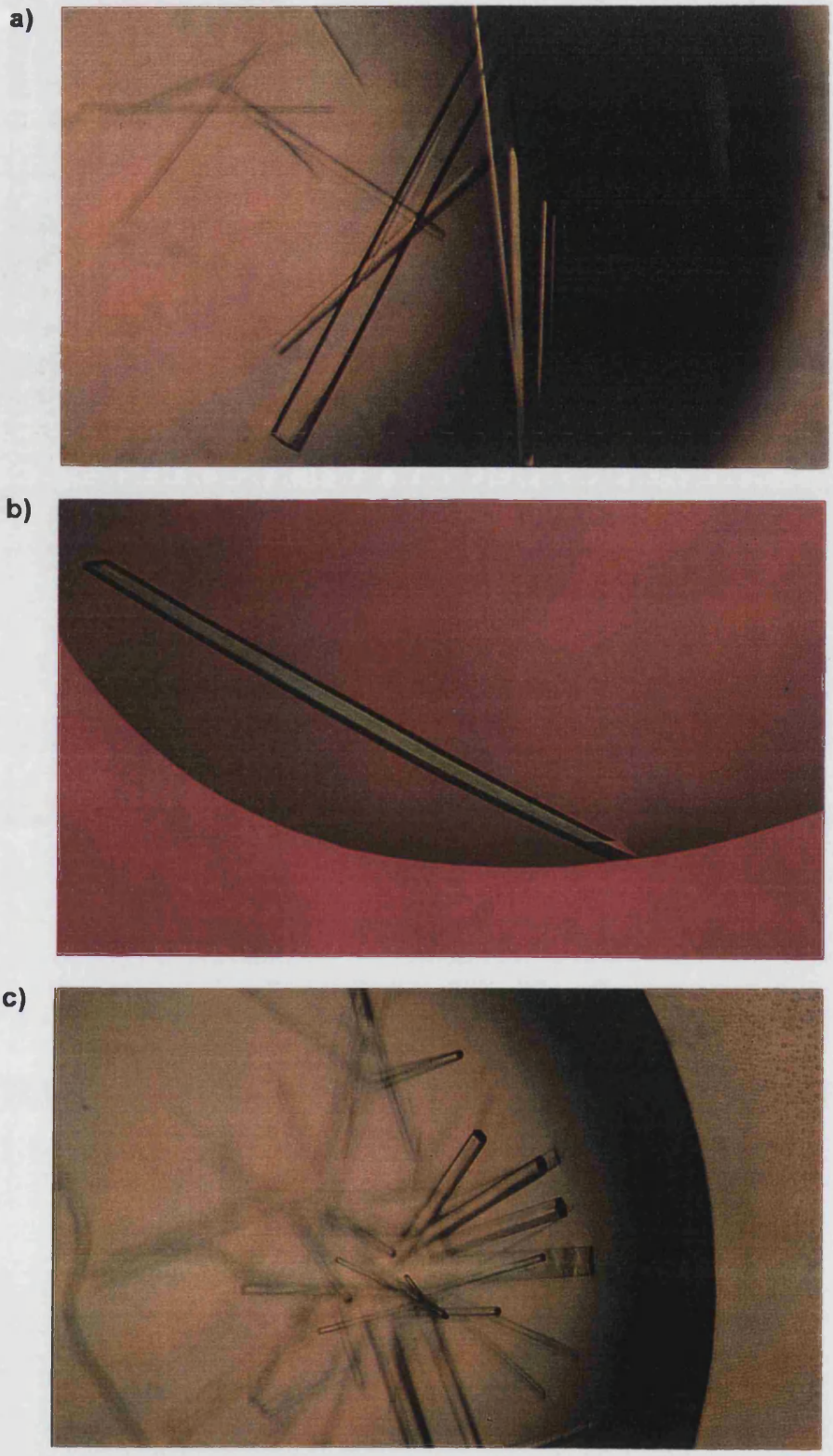


Figure 2.6: *Pw*TIM crystals a) monoclinic, b) orthorhombic, c) hexagonal

'polishing' step in the purification process to remove the remaining contaminants although in practice this was not found to improve the crystallisation and reduced the final yield.

With the recombinant *Pw*TIM purified, crystals were initially obtained from sparse matrix screening with the enzyme. Inability to improve these native TIM crystals in order to collect X-ray data prompted co-crystallisation with the two substrate/transition state analogues; 2-phosphoglycolic acid and 2-carboxyethylphosphonic acid and similar crystallisation conditions yielded crystals in each case (0.1M sodium acetate pH 4.0-4.2, 5-7% PEG 4K). The data collected from crystals of *P. woesei* TIM are summarised in the following chapter.

CHAPTER 3

P. woesei TIM; Data Collection and Crystallographic Data

3.1 Data Collection

All X-ray data were collected 'in house' on a MarResearch image plate (180mm and 300mm) mounted on an Enraf-Nonius Cu rotating anode X-ray source operating at 45kV and 80mA (Cu K_α source λ=1.542Å). The crystal was oscillated in a stepwise fashion, typically by an angle (ϕ) of 0.5° about an axis perpendicular to the X-ray beam with an exposure time varying between 5-10 minutes for contiguous frames.

Initial data were collected at room temperature, with crystals being mounted in quartz capillary tubes, and mother liquor being drained before sealing the tubes with beeswax. Due to the deterioration of the crystals in the X-ray beam during data collection experiments, the widely used method of cryo-cooling crystals to a temperature of around 100K using liquid nitrogen was carried out to enable the collection of complete data sets. In order to do this, the mother liquor was 'cryo-protected' with the addition of 35% glycerol so that the disordered water in and around the crystal would freeze amorphously and thus not contribute to the diffraction pattern. Due to the fragile nature of the *P. woesei* TIM crystals, sequential addition of cryo-protectant was necessary to avoid cracking. Glycerol was added in 10% steps, allowing 10 minutes equilibration in both 10% and 20% solutions, 5 minutes in 30%, with 2 minutes in 35% glycerol before 'flash freezing'.

3.2 Data Processing

The data were processed using the DENZO/SCALEPACK software [Otwinowski & Minor, 1997]. The various steps in data reduction have been discussed in detail [Sawyer *et al*, 1993]. DENZO reduces the raw data to a list of reflection indices and corresponding measured intensities. It also accounts

for corrections due to non-uniformity of the detector and backstop positioning. Auto-indexing of first image(s) was carried out by identifying a subset of 'bright' spots to determine crystal orientation, and identify space group, with subsequent refinement of cell and detector parameters (with importance on proper centring of the X-ray beam). The next step was prediction of expected reflection positions from the refined parameters and integration of reflection intensities for the following frames. The data were then reduced to a unique set of reflections for the specified space group. Scaling of data and output of statistics with 'post refinement' of cell and detector parameters allowing iterative reintegration of intensities adjusting such factors as mosaicity and spot size to improve data statistics. The position of screw axes were also determined from systematically absent reflections. The CCP4 program **TRUNCATE** [French & Wilson, 1978] was then used to produce a final output of Structure Factors and corresponding standard deviations.

3.3 Precision and Resolution Limits

The data quality were assessed in **SCALEPACK** by the merging R value on reflection intensity (I)

where $$R_{merge} = \frac{\sum_{hkl} |I(i) - \langle I \rangle|}{\sum_{hkl} I(i)} \times 100\%$$

representing the sum of the deviation of absolute I measurements (where *I(i)* is the *i*th measurement of the intensity of the reflection) from the mean value $\langle I \rangle$ over the sum of all I measurements. The cut off for R_{merge} at higher resolution was taken as approximately 25%. Significance of data were established by examining statistics describing signal-to-noise. Data were considered significant when 50% of reflections had an $I/\sigma(I)$ value >3 for a given resolution shell.

3.4 Data

Crystals of TIM complexed with 2-PG were found to be predominantly of an orthorhombic space group ($P2_12_12$) whilst TIM complexed with 2-CP were found to crystallise in a monoclinic space group ($P2_1$). The presence of screw axes was indicated by the systematically absent axial reflections (e.g. for $P2_1$ 0k0 are only present when k=2n). The unit cell parameters and data statistics are summarised in the table below (figure 3.1). In addition, crystals of a hexagonal space group appear to also cohabit the same drops as the orthorhombic crystals but when exposed to the X-ray beam were found to diffract poorly to a resolution limit of 6Å. For this reason data have not been collected for the hexagonal crystal form. As has been mentioned earlier, collection of data at 100K using liquid nitrogen enabled improved completeness of data sets for the orthorhombic and monoclinic crystal forms. It should be noted that freezing of the crystals gave an expected shortening in cell dimensions of around 0.5% due to ordering of solvent in the crystal. The monoclinic crystals diffracted to a slightly higher resolution (2.6Å) than that of the orthorhombic form. The orthorhombic data collected at 100K (O1_F) extended to 2.9Å but due to the formation of ice rings and the fact that the data was very weak at higher resolution the data could only be processed to 4.0Å. It is possible that the orthorhombic and monoclinic crystal forms can be easily interconverted on freezing; on a single occasion when freezing what was expected to be an orthorhombic crystal due to the fact that it had been crystallised in the presence of 2-PG (M2_F) was actually found to be monoclinic with beta $\cong 90^\circ$. Examination of the self rotation function gives further insight into this situation and proposes the close relationship between the two crystal forms.

Data Set & Temperature of Collection *	Space Group	Cell Dimensions /Å & °	Res. Limit /Å	Observed (unique) reflections	Completeness	R_{merge}	I/sigmal	Top Shell /Å
O1_RT (2pg_7) 2-PG, RT	$P2_12_12$	a=90.01 b=156.84 c=80.20	3.7	61526 (12713)	71.7 (74.2)	8.7 (13.9)	8.3 (4.5)	3.83-3.70
O1_F (4mar_1) 2-PG, 100K	$P2_12_12$	a=89.42 b=155.92 c=79.47	4.0	95933 (9907)	98.8 (97.7)	11.0 (14.0)	10.4 (8.8)	4.14-4.00
M1_RT (24jun_8) 2- CP, RT	$P2_1$	a=80.20 b=90.30 c=147.24 β=93.18	2.9	216801 (46988)	91.8 (89.6)	8.3 (24.3)	10.6 (3.1)	3.00-2.90
M1_F (7mar) 2-CP, 100K	$P2_1$	a=79.07 b=89.19 c=145.39 β=92.80	2.6	870586 (60910)	98.1 (90.0)	10.0 (26.2)	17.1 (5.5)	2.71-2.62
M2_F (21Nov_1) 2-PG, 100K	$P2_1$	a=88.24 b=152.06 c=78.9 β=89.92	3.0	224206 (41686)	96.7 (95.5)	8.5 (20.5)	9.1 (3.8)	3.14-3.00

Figure 3.1: Table summarising native data sets collected for the *Pw*TIM. Values are defined for all reflections with those in parentheses corresponding to the top resolution shell (as defined in the final column). RT= Room temperature. *Disk file names are also give in parentheses.

3.5 Solvent Content

The percentage solvent content of the protein crystals (for both space groups) was calculated from the approximation below [Mathews, 1968], in order to predict the number of tetramers in the asymmetric unit

$$V_{solvent} \cong (1 - 1.23/V_m) \times 100\ \%$$

where V_m = Crystal volume per unit of protein molecular weight

In this case the *P.woesei* TIM monomer M.W.=23 657 Da. Calculation for the orthorhombic space group ($P2_12_12$) and assuming 1 tetramer in the asymmetric unit, gave a value for V_m of 2.92Å^{3}Da^{-1}, corresponding to a solvent content $V_{solvent}$=58%, implying that there is 1 tetramer in the asymmetric unit for this space group. Calculation for the monoclinic space group ($P2_1$) gave a V_m of 2.71Å^{3}Da^{-1} and $V_{solvent}$=55% for two tetramers in the asymmetric unit.

3.6 Non-Crystallographic Symmetry (NCS)

A homo-tetramer can be arranged such that it will exhibit one of two types of symmetry. The most commonly observed is to have three perpendicular two-fold symmetry axes; 222 symmetry, as seen in many tetrameric proteins, such as the glycolytic enzyme glyceraldehyde-3-phosphate dehydrogense [Komdorfer *et al*, 1995], or four-fold symmetry, as observed in influenza virus neuraminidase [Varghese *et al*, 1983]. As the known Eukaryal and Bacterial TIMs are dimers, it seems more likely that the tetrameric *Pw*TIM will exhibit 222 NCS. This supposition was supported by information gained by calculating the self rotation function. This was done using the program GLRF [Tong and Rossmann, 1997]. The self rotation function in the program GLRF is represented by a spherical polar system where angles ϕ and ψ define a rotation axis and κ is the rotation about this axis. ϕ is the angle from the x-axis in the xy plane and ψ the angle from the z-axis. This can also be plotted as a stereographic projection (shown for the orthorhombic (O1) data in figure 3.5 and monoclinic (M1) data in figure 3.6) for the cross-section κ=180° , and

peaks in this section will correspond to two-fold NCS axes of symmetry. The orthogonalisation is such that in the monoclinic space group the unique b-axis has been made the polar axis, and related axes of similar length in the orthorhombic crystal form have been chosen such that they are in the same orientation (the orthogonalisation is shown for crystal forms M1 and O1 in figure 3.2 below). The presence and orientation of the three orthogonal 2-fold axes in the stereographic projection were observed for both the $P2_1$ and $P2_12_12$ data being indicative of 222 NCS. Furthermore, no peaks were present in the cross section for $\kappa=90^{\circ}$, which would correspond to four-fold symmetry. The angles between the peaks (θ) are calculated from direction cosines.

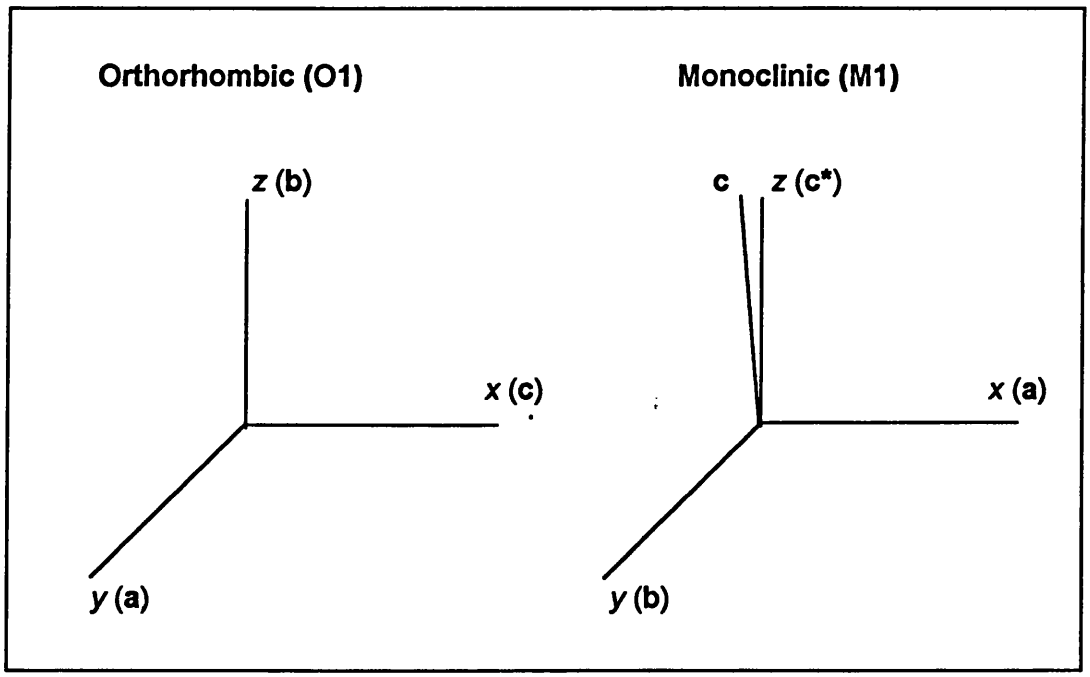


Figure 3.2: The orthogonalisation for the orthorhombic (O1) and monoclinic (M1) space groups

3.6.1 Orthorhombic ($P2_12_12$) data

Data used in the calculation of the self rotation function were from 10-4Å and a 20Å Patterson sphere was used. In this case the angle ϕ is the angle between the x-axis (c) in the xy plane and ψ is the angle from the z-axis (b). Peaks A1,A2,A3 form a 222 NCS set. Interactions of this set (A1,A2,A3) with the crystallographic 2-fold axes give rise to other related 222 NCS sets which are labelled B,C and D. Peaks labelled X (being 36% of the origin height) arise from interactions between the crystallographic screw axis along y (a) and the NCS 2-folds A3, B3, C3 and D3.

Peak no.	ϕ°	ψ°	% of origin peak
A1	81	69	33
A2	135	147	22
A3	180	66	36

Figure 3.3: table containing data of peaks seen in the stereographic projection (figure 3.5) for the orthorhombic space group ($P2_12_12$).

3.6.2 Monoclinic ($P2_1$) data (M1)

Data used in the calculation of the self rotation function were from 10-4Å and again a 20Å Patterson sphere was used. In this case the angle ϕ is the angle between the x-axis (a) in the xy plane and ψ is the angle from the z-axis (c*). In this case peaks A1, A2 and A3 also form a 222 NCS set. The second set involving peaks B1,B2 and B3 are crystallographically related via the 2_1 screw axis (simply a two-fold in Patterson space) along the unique b-axis along y (b). There are two further peaks, which have been labelled X and also arise from the interaction of the NCS two-folds A3 and B3 with the 2_1 axis.

Peak no.	ϕ°	ψ°	% of origin peak
A1	78	69	48
A2	135	144	44
A3	180	63	79

Figure 3.4: table containing data of peaks seen in the stereographic projection (figure 3.6) for the monoclinic space group ($P2_1$).

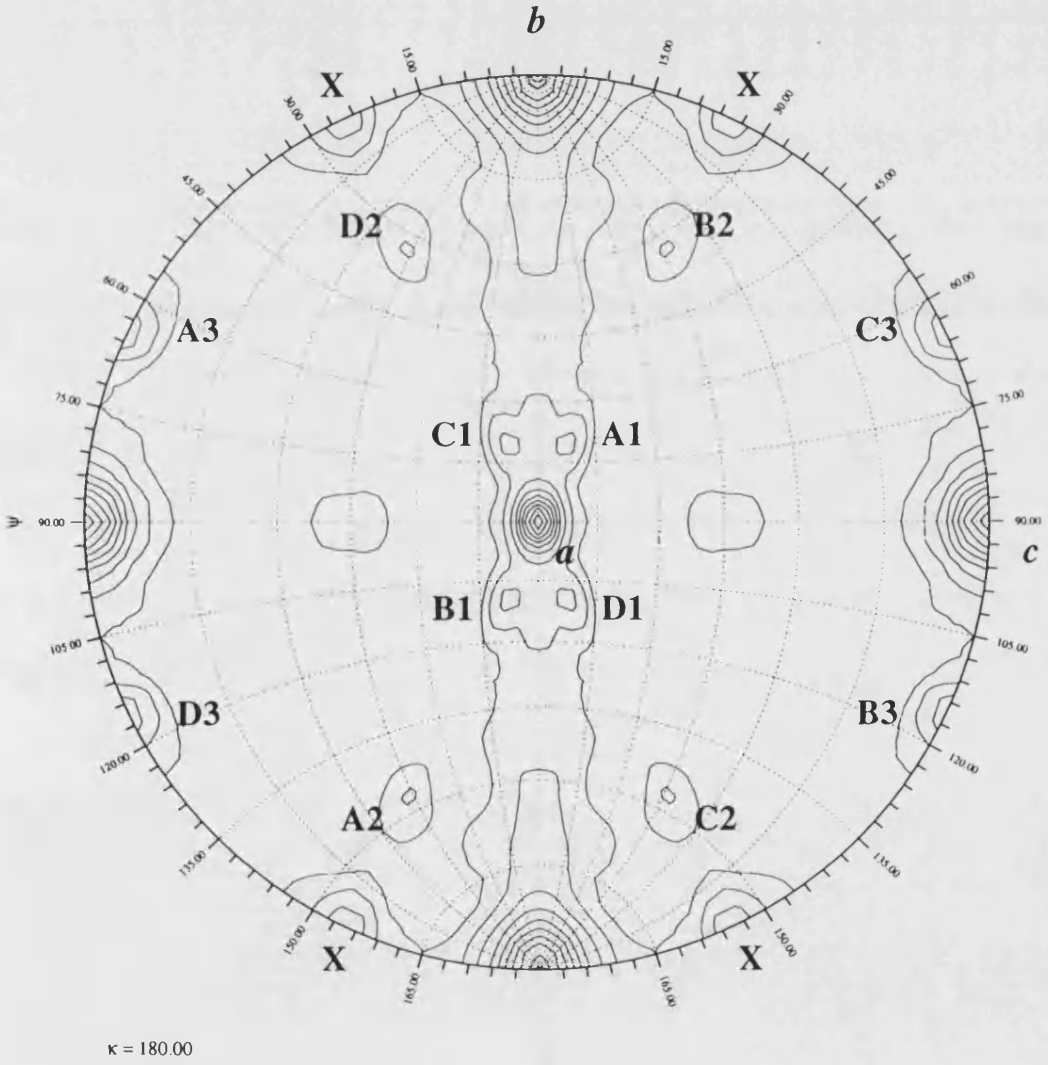


Figure 3.5: Stereographic Projection of the Self Rotation Function (κ = 180°) for the orthorhombic space group (O1) using a 20Å Patterson sphere.

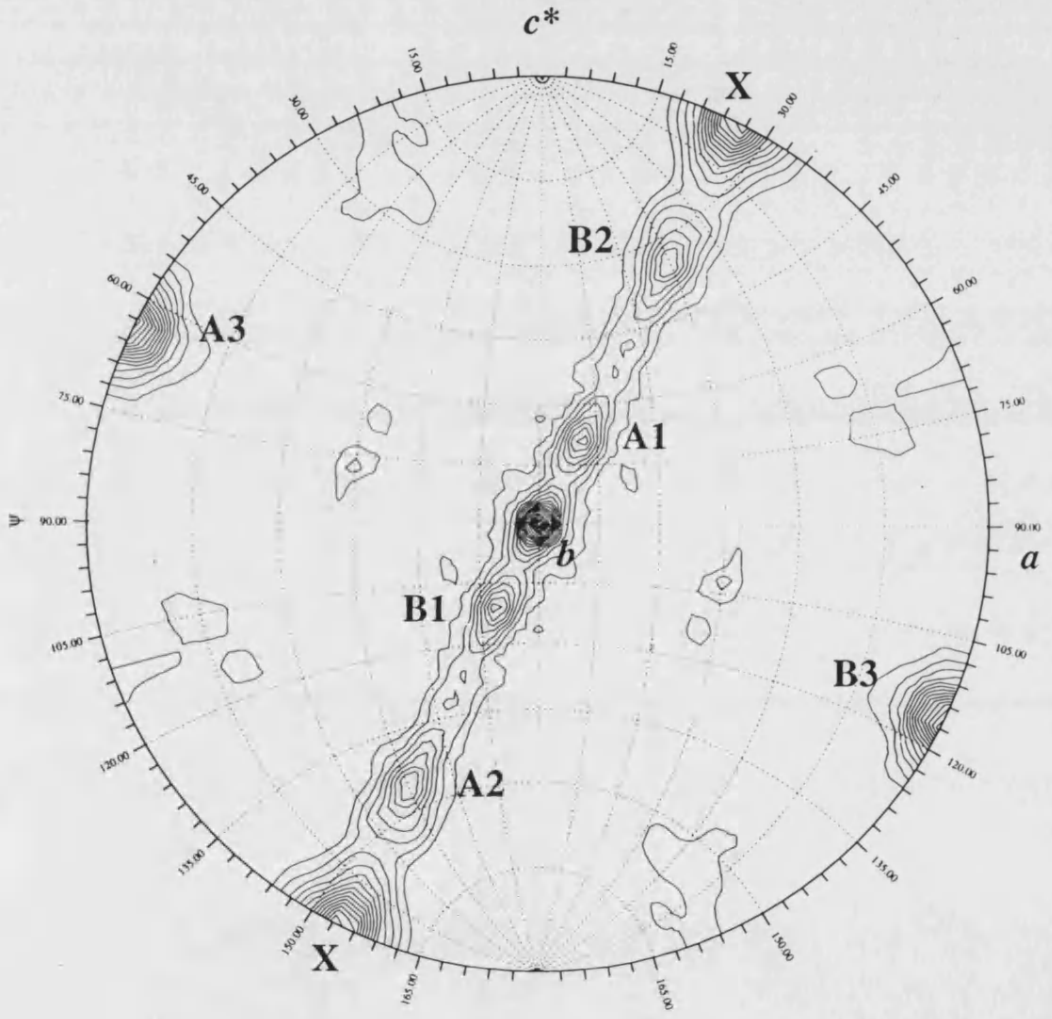


Figure 3.6: Stereographic Projection of the Self rotation Function (κ = 180°) for the monoclinic space group (M1) using a 20Å Patterson sphere.

Only one unique set of NCS axes was observed in the self rotation function for the monoclinc (M1) data set, but the solvent content calculations suggested two tetramers in the asymmetric unit. Analysis of the native Patterson for the $P2_1$ data revealed a significant peak at position of approximately 1/2, 1/4, 1/2 which was 50% of the origin peak height. This lead to the conclusion that the two tetramers in the asymmetric unit are in the same orientation and related by a simple translation. This pseudo-centring condition manifests itself in the diffraction pattern. The program HKLVIEW was used to look at data corresponding to individual planes of the reciprocal lattice and systematically weak (almost absent) reflections brought about by the pseudo-centring (figure 3.7). The structure factor equation can be used to explain these absences by giving rise to the reflection conditions below.

For the centring at x+1/2, y+1/4, z+1/2;

$$F(hkl)=\sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+ky_j+lz_j)\} + \sum_{j=1}^{n/2} f_j \exp\{2\pi i(h(x_j+1/2)+k(y_j+1/4)+l(z_j+1/2))\}$$

$$=\sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+ky_j+lz_j)\} + \sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+ky_j+lz_j)\}\exp\{\pi i(h+k/2+l)\}$$

$$=\sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+ky_j+lz_j)\}\{1+\exp\{\pi i(h+k/2+l)\}$$

where the summation *j* is over half of the atoms in the unit cell.

Therefore the reflection conditions for such a centring are

when k=2n such that k/2 is even (e.g. k=4) and h+l is odd, then (h+k/2+l) is odd and therefore $\exp\{\pi i (h+k/2+l)\}= -1 \Rightarrow$ **F(hkl) = 0**

or

when k=2n such that k/2 is odd (e.g. k=2) and h+l is even, then (h+k/2+l) is odd and therefore $\exp\{\pi i (h+k/2+l)\}= -1 \Rightarrow$ **F(hkl) = 0**

For all other cases where k=2n and h+l are such that (h+k/2+l) is even then $\exp\{\pi i (h+k/2+l)\}= +1$ and therefore **$F(hkl) = 2\Sigma f_j \exp\{2\pi i(hxj+kyj+lzj)\}$**

The above reflections where F(hkl) = 0 would therefore be absent if the tetramers were related by exactly x+1/2,y+1/4,z+1/2. As has been stated above, as this centring is only approximate, they appear as very weak reflections. In addition the true centring would in theory give a peak in the native Patterson the same height as that of the origin.

Comparison of data corresponding to the h0l, h1l, h2l and h3l zones of reciprocal space shows the effect on the diffraction pattern (figure 3.7)

3.6.3 Monoclinic ($P2_1$) data (M2)

The self rotation function (section κ=180°) for the second monoclinic crystal form (M2) (with orthogonalisation such that equivalent axial lengths are in the same orientation as for the other two space groups) can be seen in figure 3.8. As this crystal was obtained in the presence of 2-PG it was expected to be of an orthorhombic space group. The data however would only merge well for $P2_1$ (R_{merge} for all reflections of 30% for $P2_12_12$) although peaks are present that would suggest P222 crystallographic symmetry axes can be seen in the self rotation function. If $P2_1$, solvent content would suggest two tetramers in the asymmetric unit, although no significant peak is seen in the native Patterson confirming that the two tetramers must therefore be in different orientation. The similarity between this self rotation function and that calulated for data O1 can be seen and therefore given the cell dimensions (with long axis length in between that of O1 and M1 space groups, and β almost 90°) it is suggested that this data set may represent an interconversion of the two space groups upon crystal freezing.

$$F(h0l)=\sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+lz_j)\} + \sum_{j=1}^{n/2} f_j \exp\{2\pi i(h(x_j+1/2)+l(z_j+1/2))\}$$

$$=\sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+lz_j)\} + \sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+lz_j)\}\exp\{\pi i(h+l)\}$$

$$=\sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+lz_j)\}\{1+\exp\{\pi i(h+l)\}$$

when (h+l) is odd then $\exp\{\pi i(h+l)\} = -1$ therefore $F(h0l) = 0$

when (h+l) is even then $\exp\{\pi i(h+l)\} = +1$ therefore $F(h0l) = 2\sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+lz_j)\}$

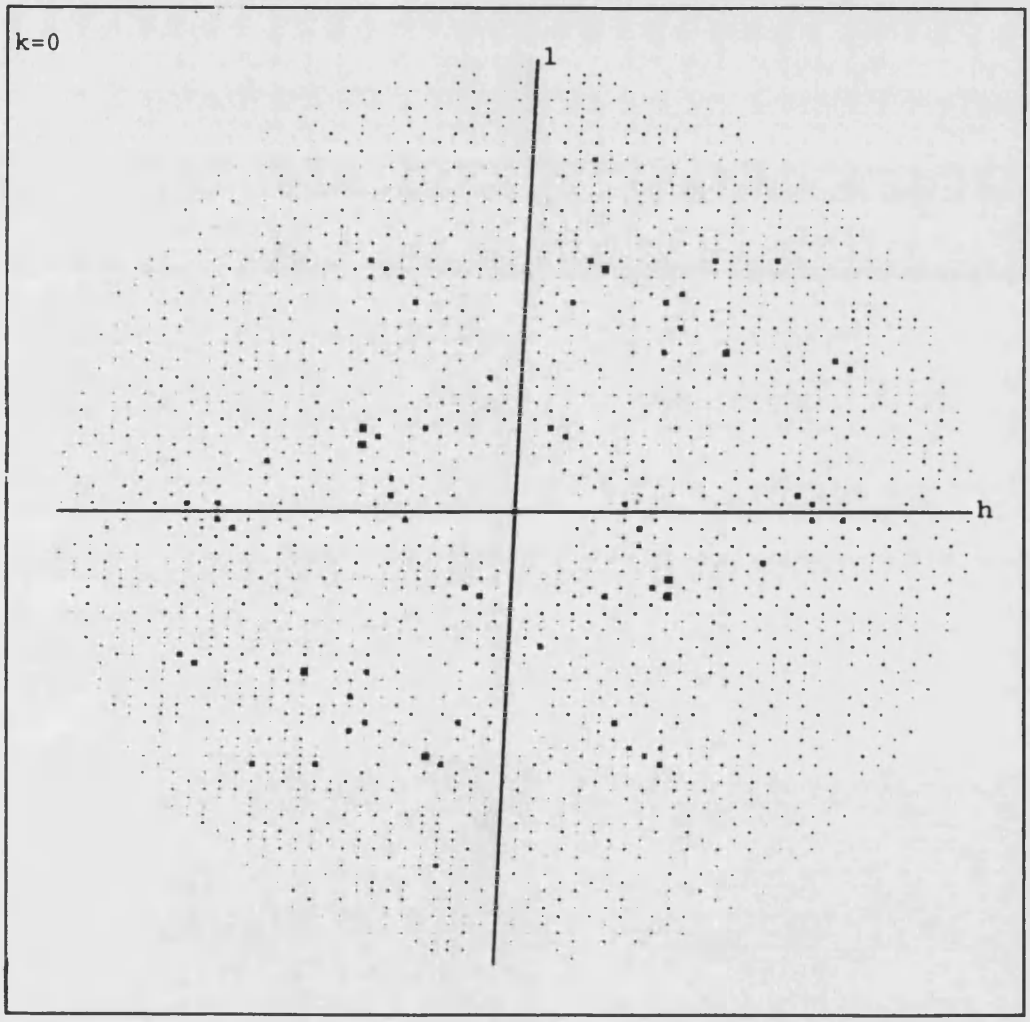


Figure 3.7a: h0l zone of reciprocal space of the *Pw*TIM monoclinic data (M1) showing systematically weak reflections when (h+l) is odd (displayed with **HKLVIEW**).

$$F(h1l) = \sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+y_j+lz_j)\} + \sum_{j=1}^{n/2} f_j \exp\{2\pi i(h(x_j+1/2)+(y_j+1/4)+l(z_j+1/2))\}$$

$$= \sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+y_j+lz_j)\} + \sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+y_j+lz_j)\}\exp\{\pi i(h+1/2+l)\}$$

$$= \sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+y_j+lz_j)\}\{1+\exp\{\pi i(h+1/2+l)\}$$

For all h and l, $\exp\{\pi i(h+1/2+l)\} \neq -1$ in general, so there are no systematically weak or absent reflections

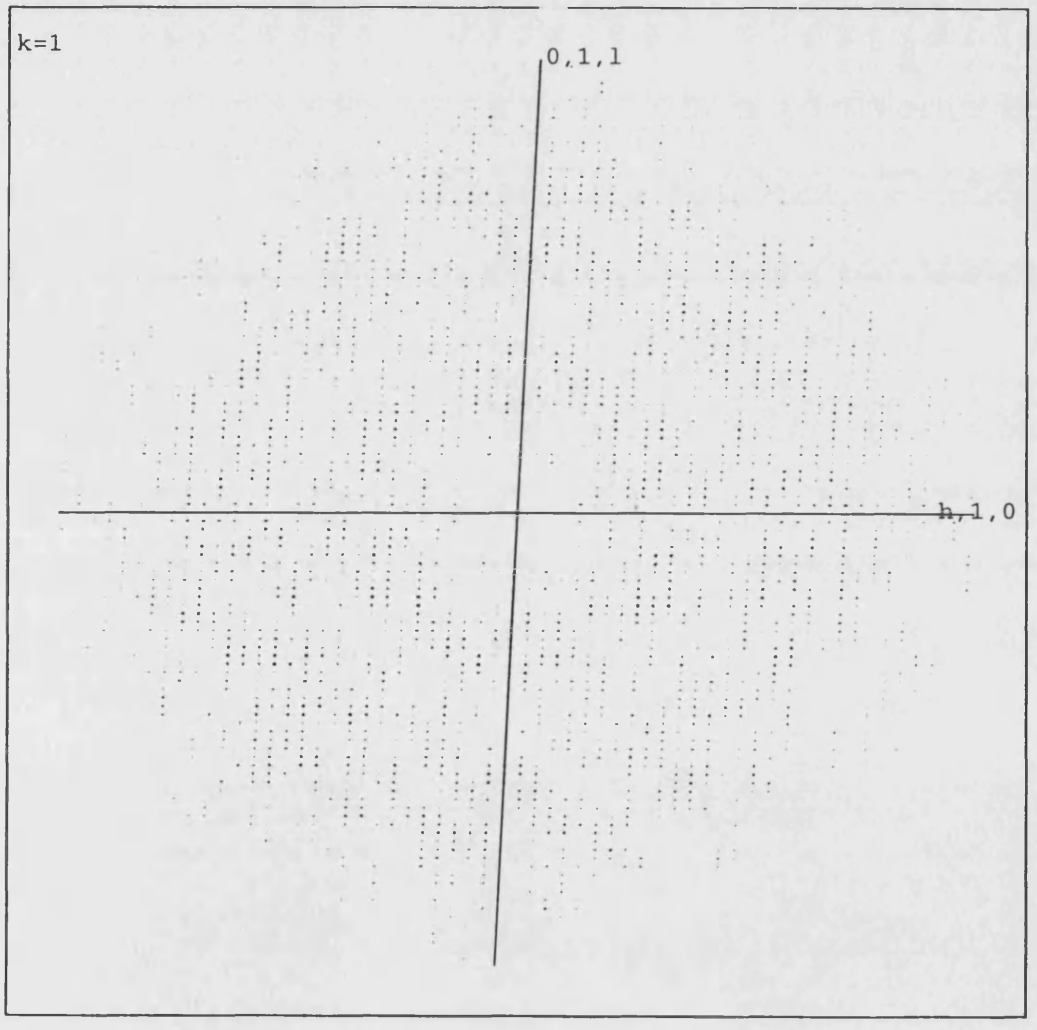


Figure 3.7b: h1l zone of reciprocal space of the *Pw*TIM monoclinic data (M1) displayed with **HKLVIEW.**

$$F(h2l)=\sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+2y_j+lz_j)\} + \sum_{j=1}^{n/2} f_j \exp\{2\pi i(h(x_j+1/2)+2(y_j+1/4)+l(z_j+1/2))\}$$

$$=\sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+2y_j+lz_j)\} + \sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+2y_j+lz_j)\}\exp\{\pi i(h+1+l)\}$$

$$=\sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+2y_j+lz_j)\}\{1+\exp\{\pi i(h+1+l)\}$$

when (h+1+l) is odd then $\exp\{\pi i(h+1+l)\}= -1$ therefore $F(h2l)= 0$ (for (h+l) even)

when (h+1+l) is even then $\exp\{\pi i(h+1+l)\}=+1$ therefore

$$F(h2l)=2\sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+2y_j+lz_j)\}$$ (for (h+l) odd)

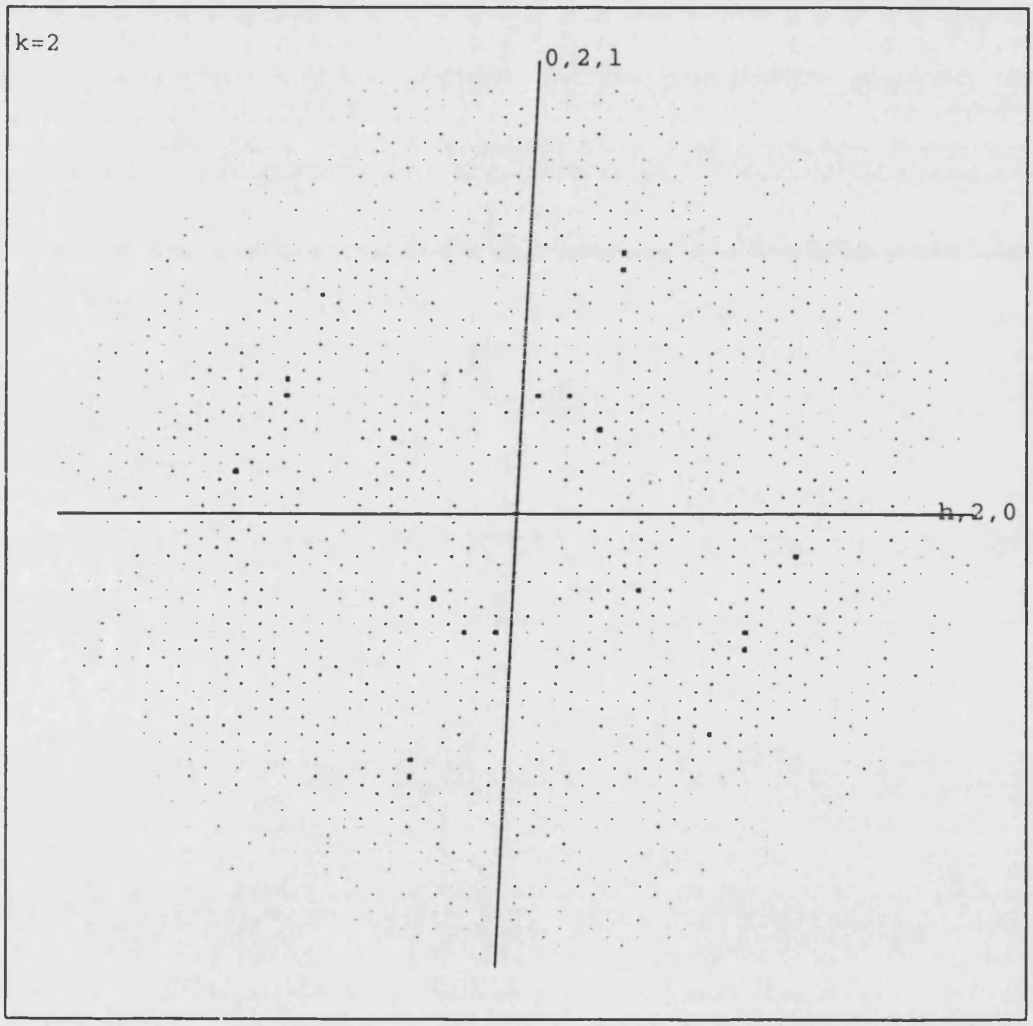


Figure 3.7c: h2l zone of reciprocal space of the *Pw*TIM monoclinic data (M1) showing systematically weak reflections when (h+l) is even (displayed with **HKLVIEW**).

$$F(h3l)=\sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+3y_j+lz_j)\} + \sum_{j=1}^{n/2} f_j \exp\{2\pi i(h(x_j+1/2)+3(y_j+1/4)+l(z_j+1/2))\}$$

$$=\sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+3y_j+lz_j)\} + \sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+3y_j+lz_j)\}\exp\{\pi i(h+3/2+l)\}$$

$$=\sum_{j=1}^{n/2} f_j \exp\{2\pi i(hx_j+3y_j+lz_j)\}\{1+\exp\{\pi i(h+3/2+l)\}$$

For all h and l, $\exp\{\pi i(h+3/2+l)\} \neq -1$ in general, so there are no systematically weak or absent reflections.

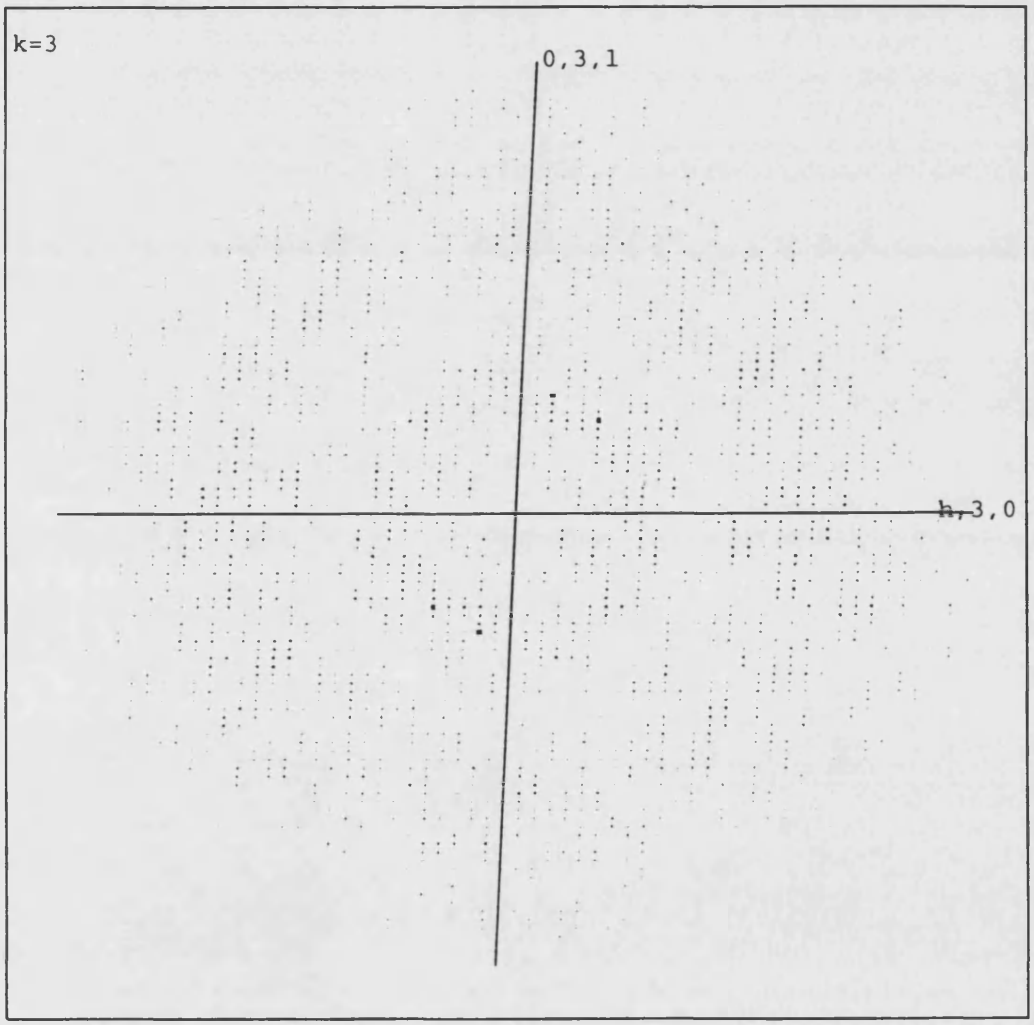


Figure 3.7d: h3l zone of reciprocal space of the *Pw*TIM monoclinic data (M1) (displayed with **HKLVIEW**).

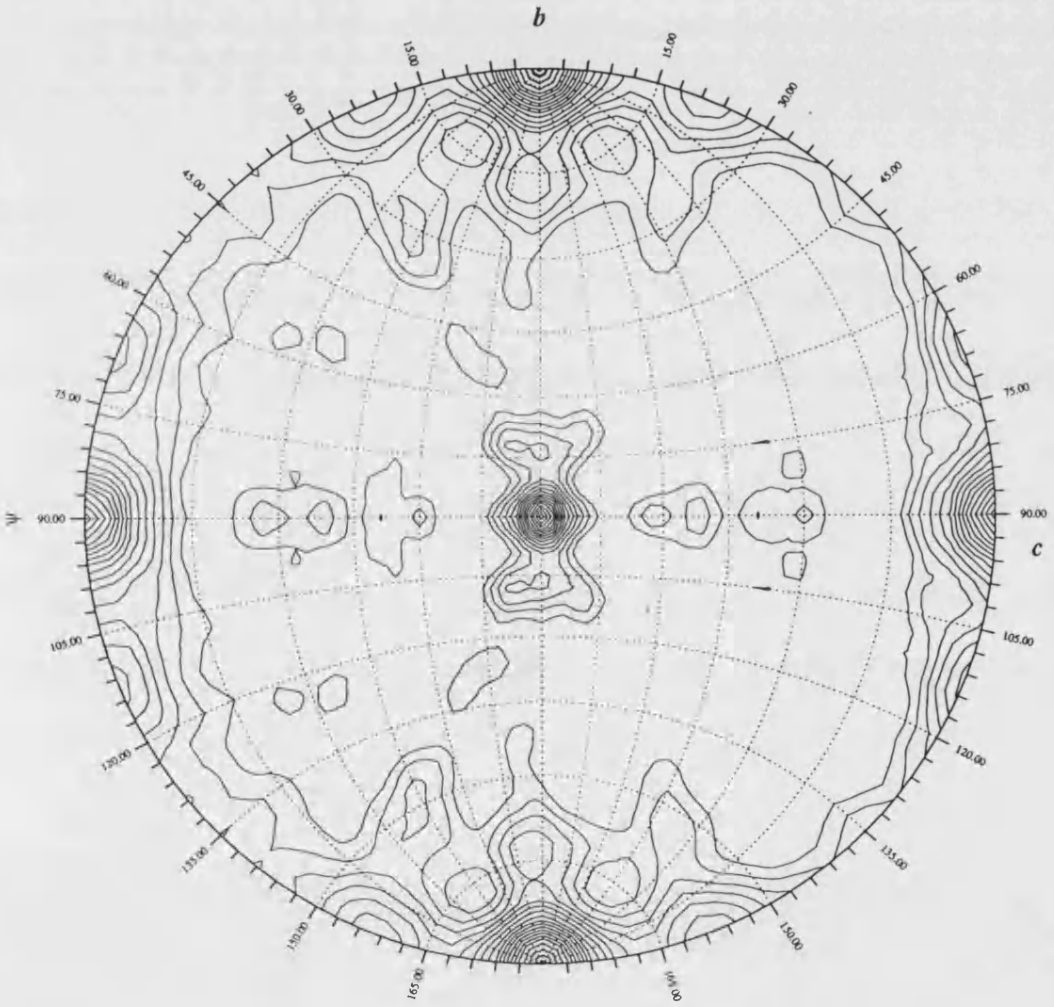


Figure 3.8: Stereographic Projection of the Self Rotation Function (κ = 180°) for the second monoclinc space group (M2) using a 20Å Patterson sphere.

3.7 Discussion

Complete data sets were collected for the *Pw*TIM/2-PG (orthorhombic, O1) and *Pw*TIM/2-CP (monoclinic, M1) complexes to a resolution of 4.0Å and 2.6Å respectively, with the hexagonal crystal form (*Pw*TIM/2-PG) showing disorder. Attempts to collect higher resolution data at the synchrotron source at EMBL (Hamburg) were unsuccessful as the crystals tried there performed no better than on the 'in house' source. Information from the self rotation function in conjunction with that obtained from the native Patterson of the monoclinic data (M1) has given insight into the non-crystallographic symmetry within the two crystals. The tetramer in the orthorhombic form is situated with one NCS 2-fold lying in the **bc** plane which corresponds to the monoclinic space group in which one of the NCS two-folds is lying in the **ac** plane. The orientation of the tetramer in both crystal forms is therefore very similar with respect to the corresponding cell axes, and this information helps explain the possible interconversion of the two space groups. It was hoped that these observations would aid the attempts in structure solution.

CHAPTER 4

P. woesei TIM; Molecular Replacement

4.1 Introduction

If the structure of a protein with a similar sequence (and hence homologous fold) is available then this may be used as an initial phasing model for the new structure; this is known as Molecular Replacement (MR). This method enables the structure to be solved using only a single native data set. The chosen phasing model must be rotated and translated to find the correct orientation (defined by 3 Eulerian angles α,β,γ [Rossmann & Blow, 1962]) and a translation (defined by a translational vector) of the model in the new unit cell, and this 6-dimensional problem is separated into rotation and translation parts [Hoppe 1957]. In this case it was hoped to use Molecular Replacement as a means of obtaining initial phases due to the high number of structures on which to base a search model, and the $(\alpha\beta)_8$-barrel of the known TIM structures being highly conserved. The following sections discuss the programs, search models and additional considerations involved in an attempt to solve the *Pw*TIM structure by this method.

4.2 Data

Molecular Replacement was attempted on all native data sets collected (see chapter 3) but efforts were concentrated on the monoclinic (M1_F) and orthorhombic (O1_F) data sets which were collected at 100K and were >98% complete. The room temperature orthorhombic data was of slightly better quality than the latter but only 72% complete. For the orthorhombic data set, solvent content implies the presence of one tetramer in the asymmetric unit. For the monoclinic data set, there are two tetramers in the asymmetric unit, however the scale of the problem is similar in both cases because of the apparent fact that the two tetramers are in the same orientation and related by a translation in the monoclinic case. The quality of the $P2_1$ data was also

considerably higher than that of the $P2_12_12$ data (with the $P2_12_12$ data set only extending to 4Å).

4.3 Search Models

As mentioned in the introduction there are TIM structures from eight organisms in the Brookhaven Protein Data Bank [Bernstein *et al*, 1977] and those used are listed in the table in figure 4.1 (the crystal structure of TIM from *Thermotoga maritma* - which is 26.2% identical at the sequence level- has also been determined but the results are yet to be published [Alvarez *et al*, 1998]). Most of the study was carried out on the first six structures in the table. At a later stage the two TIM structures which have been submitted to the PDB more recently were used for initial cross rotation and translation searches. In general the structure refined to the highest resolution for each organism were taken, and search models for Molecular Replacement were then derived from these coordinates. Although it is structural similarity that is required for success of the method, this is reflected in the sequence identity, and it is therefore an obvious assumption that *Pw*TIM may be quite structurally different from the known 3-D structures. It was assumed that the *Pw*TIM was in the closed conformation (due to the different crystallisation conditions from that of the native enzyme) therefore most of the searches were carried out using models derived from the closed structures (apart from *E. coli* and *P. falciparum* for which none were available).

It is important to note that as all known TIM structures are homo-dimeric, there is no structural data to indicate the nature of dimer-dimer interactions on formation of the tetramer. Initially searches were tried with the dimers and monomers of the six structural TIMs. The next step was the removal of the region of structure corresponding to the possible 'deletion' around helix 5 and the subsequent loop (residues 133-163 in the human TIM). Structural alignment using the COMPOSER program established the conserved core of the known TIM structures and this was used to make further adaptions to the model by pruning other loop regions, which tended to correspond to the least conserved regions of the structure. A model containing the most highly

conserved regions with respect to the sequence alignment of *Pw*TIM with the structural TIMs was also created. It has been shown that a correct MR solution can be found using a good search model which comprises only a small percentage of the scattering mass [Bernstein & Hol, 1997] and therefore several models containing only the β-strands or α-helices, or the active site region were also used. The reduction of all dimers and monomers to a polyalanine/glycine chain was also carried out due to the low sequence homology of the available search models. In addition to using single PDB files, combinations of dimers/monomers were 'overlayed' (by least squares fitting in the program O [Jones, 1978]). On superposition of monomers of the structural TIMs (and subsequently examining relative positions of the second monomer) it was seen that the dimers were of varying angles of one monomer relative to the other. A set of models with varying angles was therefore created using XPLOR [Brunger, 1992]. A model was also made in O using the sequence of *Pw*TIM based on the *T. brucei* structure.

Organism	PDB code	Res. /Å	Substrate	% Seq ID (with *Pw*TIM)	Reference
Chicken	1tph	1.8	2-PGH	20.9	Zhang *et al*, 1994
T. brucei	6tim	2.2	G-3-P	26.5	Noble *et al*, 1991
Yeast	7tim	1.9	2-PGH	20.9	Davenport *et al*, 1991
E. coli	1tre	2.6	—	22.3	Noble *et al*, 1993
B. stearo.	1btm	2.8	2-PG	20.8	Delboni *et al*, 1995
Human	1hti	2.8	2-PG	21.6	Mande *et al*, 1994
P. falciparum	1ydv	2.2	—	17.9	Velankar *et al*, 1997
V. marinus	1aw1	2.7	2-PG	23.6	Alvarez *et al*, 1998

Figure 4.1: PDB coordinates used for molecular replacement (2PG = 2-phosphoglycolate, G-3-P = glycerol-3-phosphate, 2-PGH=2-phosphogycolohydroxamate).

4.4 Cross-Rotation and Translation Searches

4.4.1 Search Parameters

The selection of several key variables must be made when using Molecular Replacement programs for the cross rotation and translation searches. Which experimental data is to be included in the calculation is obviously very important. Low resolution data (<10 Å) was not always included as it contains information mainly about solvent in the crystal. It must also be decided whether or not to include high resolution data, and this may often depend on the similarity of the search model to that of the subject protein. The size of the Patterson integration radius for the cross rotation was based on the dimensions of the protein involved in order to maximise intra-molecular vectors but exclude as many inter-molecular vectors as possible. The inter-molecular vectors were desired when calculating the translation function. The model unit cell (P1 box) must also be stipulated, and this was chosen to be a minimum size of at least that of the model plus the sphere of integration. It should be noted that although many of the parameters may be chosen intuitively based on the specific problem, a trial and error optimisation was also attempted, being aided by the speed of several of the MR programs.

4.4.2 Non-Crystallographic Symmetry

Knowledge of the positioning of NCS axes obtained from the self rotation function, can help pick out the correct solutions from the cross rotation (see 4.4.3). Another option is to compute the "locked" cross rotation function [Tong & Rossmann, 1990] which includes information about non-crystallographic symmetry in the cross rotation searches (see 4.4.5). The latter method should increase signal-to-noise in a situation where NCS is present.

With the presence of 222 NCS, for every rotation [E] that brings the monomer search model into the same orientation as one of the monomers in the crystal,

there will also be an additional set of rotations which relate this to the rest of the tetramer.

The rotation (ρ) which corresponds to the re-orientation of the model onto that of the crystal is now

$$[\rho_n] = [F] [l_n] [E]$$

where $[l_n]$ (n=1,....N) are the NCS rotation matrices. [F] is the rotation which orients the NCS matrices onto to that of the crystal (defined by the peaks in the ordinary self rotation function) and [E] is the rotation which corresponds to the orientation of the model onto one of the monomers in the crystal.

4.4.3 AMoRe

Standard cross rotation and translation searches were carried out with AMoRe [Navaza, 1994], for $P2_1$ and $P2_12_12$ data sets, with the scheme being described below. The 'sorting' step sorts hkl, Fo and σFo from the data MTZ file into and internal form. The 'tabling' routine generates structure factors for the model in a P1 box (with centre of mass at the origin). The fast cross rotation search [Crowther, 1972] was then computed in 'roting' and the translation function [Crowther & Blow, 1967] was calculated - in 'traing'- on the top solutions defined in eulerian angles. The best solution can then be fixed (and hence the origin) and translation function calculated again on other rotation solutions, to solve these relative to the first. The integral rigid body refinement 'fiting' stage was also carried out and then transformations applied to these final solutions in the 'shifting' stage in order that they correspond to the initial model. The validity of solutions could then be checked in O by analysis of the packing within the unit cell.

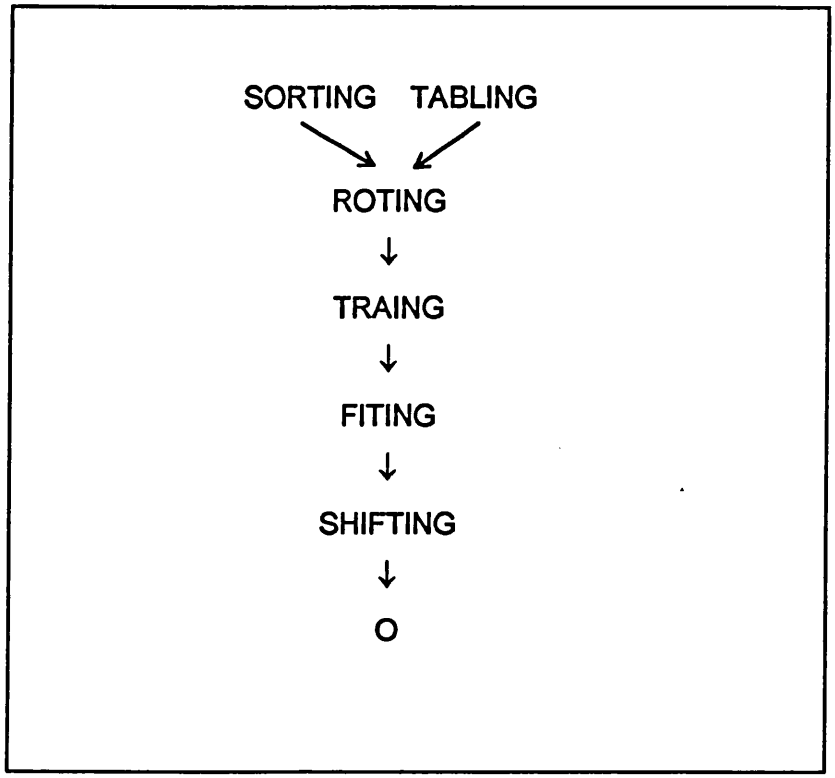


Figure 4.2: Scheme for Molecular Replacement with AMoRe.

In addition, an attempt was made to simplify the search for the $P2_1$ data by inputting information about the second tetramer in the asymmetric unit. This was achieved with the addition of two more symmetry operators; (x+1/2,y+1/4,z+1/2) and (1/2-x,y+3/4,1/2-z) to account for the second tetramer in the asymmetric unit. Thus a new space group was effectively created ($G2_1$) and this space group was added to the CCP4 space group library to allow use in AMoRe. Due to the approximate nature of the centring, it was assumed that this would only hold for low resolution.

The normalisation of structure factors (calculating E values) was carried out by dividing the data into resolution shells of constant volume and applying a scale factor so that the mean intensity does not decrease, with increasing resolution. The scaling is done by applying an exponential temperature factor to the structure factors in the program ECALC. Using E values should increase the

contrast of the Patterson map compared with using Fs. This was used in conjunction with the AMoRe program, with the T_2 translation function being calculated using the program TFFC [Tickle, 1985]. The peak resolution in the cross rotation was not improved using this method.

To incorporate NCS information, the program RFCORR [Tickle, 1998] was used to take pairs of peaks from the cross rotation function and compared these with peaks from the self rotation function (κ=180°) as calculated using POLARRFN [Kabsch, 1998]. This simplified the search by examining only those solutions which would correspond to molecules related by the two-fold NCS.

4.4.4 XPLOR

Standard cross rotation searches and translation searches were attempted using XPLOR but increased computer time (compared with AMoRe) made this less suitable for 'screening' the many different search models. One feature of the program, is that it has a Patterson Correlation-Refinement step which is run on solutions from the cross rotation search. This was thought to be useful, particularly with the dimer as a search model, as the monomers can be refined separately. The program also enabled a reduction of the rotation search by first aligning a TIM dimer along an NCS two-fold axis (and performing a rotation search about this axis).

4.4.5 Replace

REPLACE [Tong, 1993] is a molecular replacement package which in addition to ordinary rotation and translation functions, also encorporates the general locked rotation function as described in 4.4.2. An ordinary self rotation function was calculated using GLRF [Tong & Rossmann, 1990] and the NCS information (for either orthorhombic or monoclinic space groups) input into the locked cross rotation search which was calculated using the same program. The translation function was then computed using TF [Tong, 1993].

4.5 Discussion

No MR solutions were found with signal-to-noise significantly higher than the background, using any of the described methods and a large number of search models. The crystal packing of any promising solutions was analysed using the program O but all were ruled out due to severe clashes with crystallographically related models. Attempts were made to obtain a solution for all data collected; although the orthorhombic data collected was comparatively poor, any solution found for this could then have been applied to the monoclinic cell (with data extending to a higher resolution). There are several reasons that can be proposed for the failure of Molecular Replacement. Most importantly, the homology of the model was quite low (sequence identity of 20-26%) for success of the method. No knowledge about the specific interactions involved in the formation of the tetramer meant that the searches were carried out with dimers and monomers, comprising a low percentage of the asymmetric unit. It is also possible that the nature of the dimer varies from the known TIMs in formation of a tetramer, and if this is the case, a monomer is probably a better search model. The scale of the problem was also considerable; with four monomers per asymmetric unit for the orthorhombic cell and eight in the monoclinic. Knowledge from the self rotation function regarding the nature of the tetramer, the position of NCS axes and the pseudo symmetry in the M1 data failed to simplify the problem. It was therefore decided to embark on the Isomorphous Replacement method discussed in the following chapter.

CHAPTER 5

P. woesei TIM; Isomorphous Replacement

5.1 Introduction

Due to the lack of success using Molecular Replacement to obtain protein phases, the method of Isomorphous Replacement was pursued (theory discussed in appendix 2). This method involves the binding of heavy atoms to a small number of (specific) sites on the protein which cause a measurable perturbation in the diffraction intensities. It is important that the binding of heavy atoms cause no change in the unit cell parameters or rearrangements within the cell. The accepted limit for isomorphism, is about 0.5% change in cell dimensions [Crick & Magdoff, 1956].

As a result of phase ambiguity, at least two derivative data sets are generally required to solve the structure by multiple isomorphous replacement (MIR). Measurement of isomorphous differences also can be used in conjunction with Anomalous Scattering (SIRAS, MIRAS). In the case of non-crystallographic symmetry, then averaging in conjunction with the density modification procedure, can break the phase ambiguity of a Single Isomorphous Replacement (SIR).

5.2 Preparation of Heavy Atom Derivatives

5.2.1 Method

The most common methods employed to introduce heavy atoms into a target crystal (reviewed in [Blundell & Johnson, 1976]), include "soaking" the crystals in mother liquor (or suitable buffer) containing the chosen heavy atom salt or co-crystallisation in the presence of the reagent. The absence of bound metal ions in the TIM structure ruled out the commonly used method of replacement with metals of a similar radius. Inhibitors or co-factors incorporating heavy

atoms have also been used successfully [Amone *et al*, 1971] but were not available in this case. Another option is to modify the sequence by mutagenesis such as the creation of a surface cysteine residue but this has not been attempted. The heavy atom soaking method was used initially as co-crystallisation may interfere with interactions involved in the crystallisation process leading to greater chance of non-isomorphism.

The rationale behind selection of suitable derivatives can be made with respect to the presence of particular amino acids but solvent accessibility could only be roughly estimated from the low homology model (even if a side-chain is not buried in the core, further complications could arise from involvement in crystal contacts). Protein ligands are often divided into two classes; hard and soft [Pearson, 1963]. Hard ligands such as the carboxylate groups of glutamate and aspartate (and C-terminus) and hydroxyl groups of serine and threonine tending to prefer the electronegative class A metal ions (e.g. uranyl) involved in electrostatic interactions. Soft ligands such as the sulphur groups of methionine and cysteine and the imidazole of histidine tend to favour more covalent interactions with class B metal ions such as Pt, Pd, Au or Hg. In the case of the *Pw*TIM, there are three cysteine groups in the amino-acid sequence, however analysis of a model based on the *T. brucei* TIM structure suggests that these are not likely to be solvent accessible. Despite these classifications, many interactions are of a less specific nature and this diversity of chemistry leads to reaction times varying from seconds to months. Additional considerations such as differing reactivity of complexes, pH, heavy atom salt concentrations, temperature and steric effects result in the method containing a large element of trial and error. In an attempt to find suitable derivatives other atoms were used such xenon gas as it is thought to occupy hydrophobic 'pockets' within the protein and iodine which has been shown to interact with the pi electron clouds of tyrosine residues as a I^{3-} or I^- species. As a better native data set was available for the monoclinic space group (that crystallised with 2-CP) it was thought they were more suitable for heavy atom experiments.

5.2.2 Heavy Atom Soaks

Due to the fact that most groups are reactive at higher pH due to the lower likelihood of, for example, histidine or cysteine residues existing in a protonated form, a condition was found at which the crystals were stable but would be more likely to react than that at which they were grown. 0.1M MES buffer pH 6.5 (6% PEG 4K) was used for these purposes (at a higher pH than this the crystals appeared to be unstable over a period of hours). Salts of heavy atoms were dissolved in the above buffer and the *Pw*TIM crystals transferred to the solution and stored in a capillary tube. The crystals appeared to be very unstable in most of the heavy atom solutions tried at room temperature. Stability was shown to increase on setting up these soaking experiments at 5°c, and therefore the bulk of experiments were carried out at this temperature. In the case of Xenon, the crystal was placed in a hair loop, situated in a sealed container (at room temperature) which was filled with xenon gas for 10 minutes [Soltis *et al*, 1997] immediately prior to data collection.

5.2.3 Heavy Atom Co-crystallisation

All heavy atom co-crystallisation trials were set up using conditions similar to those optimised for growth of the *Pw*TIM/2-CP crystals being 0.1M NaAc pH 4.5 and 6% PEG 4K containing 20mM 2-CP. All the heavy atom reagents were added to the crystallisation buffer at a concentration of 2mM (1mM final concentration in hanging drop). Most of the crystals appeared between 1 day and 1 week.

In addition an attempt to co-crystallise *Pw*TIM with Na_2WO_4 (in the absence of inhibitors) as it was thought that tungstate may bind to the phosphate binding site. The fifty conditions of the Hampton Research Crystal Screen I were set up with 50mM Na_2WO_4. After 3 days conditions 15, 18, 40-43 and 49 all yielded small needles. Unfortunately the attempts to improve these crystals were unsuccessful. The TIM/Na_2WO_4 50mM was also used to set up a matrix

Compound	Conc /mM	pH	Temp /°C	Time	Comments
$Hg(II)NO_3.H_2O$	1	6.5	5	30 mins	cracked immediately
$CH_3Hg(II)Cl$	1	4.5	5	2days	cracked
PIP	1	6.5	5	2 days	cracked
PCMB	0.5	6.5	5	30 mins	cracked
PCMBS	0.1	6.5	5	3 days	data collected - disordered crystal
$K_2Pt(IV)Cl_6$	1	6.5	5	24hours	data **PTCL6_2** (after 2 days crystal cracked under these conditions)
$K_2Pt(II)Cl_4$	1	6.5	5	2 days	data **PTCL4_4**
Pt(II)TCl (T=tepyridine)	1	6.5	5	2 days	cracked
$K_2Pt(II)(NO_3)_4$	1	4.5	5	2 days	data **PT2_1**
$PbAc_2$	2	6.5	5	30 mins	cracked
$K_3UO_2F_5$	1	6.5	5	15 hours	data collected - disordered crystal (after two days crystal cracked)
$AgNO_3$	1	6.5	5	10 mins	cracked
$CdCl_2.2.5H_2O$	2.5	6.5	5	4 days	data **CD_6**
$BaCl_2.2H_2O$	2	6.5	5	12 hours	cracked
$Ho(III)Cl_3.6H_2O$	1	6.5	5	12 hours	cracked
$SmCl_3$	1	6.5	5	10 mins	cracked
$Ce(SO_4)_2$	1	6.5	-	-	would not dissolve
LaCl	1	6.5	5	12 hours	cracked
$Yb_2(SO_4)_3$	1	6.5	-	-	would not dissolve
$Th(NO_3)_4$	1	6.5	5	10 mins	cracked
$TO_6\ Br_{14}$	0.1	6.5	5	12 hours	cracked
KI/I_2	-	6.5	5	1& 7 days	data **I3_3** and **I3_4** respectively
Xe(g)	-	-	-	10 mins	data **XE_2**

Figure 5.1: table summarising soaking experiments (data file names in bold type).

around the condition used for TIM/2PG and TIM/2CP complexes but all trials showed some degree of precipitation and no crystals were observed.

5.2.4 Seleno-Methionine *Pw*TIM

As the search to obtain heavy atom derivatives was proving difficult, an attempt was made to prepare Seleno-Methionine *Pw*TIM in order that a Multiwavelength Anomalous Dispersion (MAD) experiment [Hendrickson, 1991] could be carried out using the variable wavelength X-ray source at EMBL, Hamburg. The *Pw*TIM has five methionine residues per monomer (twenty per asymmetric unit in the orthorhombic form and forty per asymmetric unit in the monoclinic form)

The SeMetTIM was expressed and purified in Essen, Germany (Reinhard Hensel, unpublished work). On receiving the protein, it was necessary to dialyse with the buffer previously used for crystallisation trials (0.1M TRIS/HCl, pH8.5, 25mM NaCl, 2mM EDTA). As it has been reported that seleno-methionine proteins tend to crystallise in conditions very similar to that of the native protein [Hendrickson & Ogata, 1997] this was assumed in setting up crystallisation trials with the *Pw*TIM. The protein was concentrated to several final concentrations of between 5 and 20mg/ml and crystallisation trials set up by the hanging drop method (in the presence of 20mM 2-CP or 2-PG) in a matrix around the condition known to yield crystals for the native *Pw*TIM (chapter 2). It was thought that this would allow for any difference in the solubility of the SeMetTIM from that of the native protein. Unfortunately, these crystallisation trials were unsuccessful and no sign of crystallisation was observed. Some of the higher protein/precipitant concentration samples were shown to precipitate out of solution. We were therefore unable to measure X-ray data on the SeMetTIM.

Compound	Comments/Data Collection
CH_3HgCl	No crystals
$Hg(II)Ac_2$	No crystals
$Hg(II)Cl_2$	Small rods- Poor diffraction (<10Å)
$Hg(II)NO_3.H2O$	No crystals
$Hg(II)(SCN)_2$	Crystals appearing after 2 days appeared to be of different morphology (possibly tetragonal)- diffraction <8Å
$K_2Pt(IV)Cl_6$	Small rods
$K_2Pt(II)(CN)_4.3H_2O$	No crystals
$K_2Pt(II)Cl_4$	Heavy precipitate
Pt(II)TCl (T=tepyridine)	No crystals
$K_2Pt(II)(NO_3)_4$	Small rods
$CdCl_2.2.5H_2O$	Needles
$BaCl_2.2H_2O$	Small rods obtained after 3 days - some data collected to low resolution but crystal disordered.
$PbAc_2$	Needles
$NaAuCl_4$	No crystals
Na_2WO_4	No crystals
UO_2Ac_2	No crystals
UO_2SO_4	Precipitate
$SmCl_3$	Small rods -data collected but no diffraction >10Å
$LaAc_3$	Initial data collected to 6Å but showed crystal disorder.
$Th(NO_3)_4$	small needles
Na_3IrCl_6	Rod like crystals obtained after several days data set IR_1
$Yb_2(SO_4)_3$	No crystals

Figure 5.2: table summarising co-crystallisation experiments (data file name in bold type).

5.3 Analysis of Heavy Atom Data

After collection of the putative derivative data set, this must be scaled to the the native data (assuming isomorphism) before calculation of isomorphous difference Pattersons to locate the heavy atom sites. The heavy atom parameters can then be refined to allow subsequent calculation of protein phases. The calculation of difference Fouriers can then be used to check for other heavy atom sites and also to find sites for other derivatives. This process (and the programs used) is summarised in the diagram in figure 5.3 and described in more detail below.

After initial column merging of data with **MTZUTILS** [Dodson & Terry, 1998], a scale factor was determined to scale the derivative to the native data set using **SCALEIT** [Evans *et al*, 1998] (which was used to apply an anisotropic temperature factor). The Kraut procedure in the program **FHSCAL** [Tickle, 1998] was also used as a precursor to using **VECREF** [Tickle, 1998].

An Isomorphous Difference Patterson using Fourier coefficients of $(k|F_{PH}|-|F_P|)^2$ could then be calculated using **FFT** [Ten Eyck, 1998] (in which there will be vectors corresponding to all pairs of heavy atoms in the unit cell). Vectors between one atom and its general equivalent positions (related by the space group symmetry) are called Harker vectors. In the $P2_1$ case described here, for a heavy atom with coordinates (x,y,z), there is a symmetry related position at (-x,1/2+y,-z) giving a vector at 2x,1/2,2z. For this reason, in the Patterson (u,v,w) map, the 'Harker section' is v=1/2. This section could then be plotted and examined using the program **NPO**.

Patterson search programs (**VECSUM** or **SHELX** [Sheldrick, 1991]) were used to find heavy atom sites. Once the first site is derived from the Harker vectors, the origin is fixed, (for the monoclinic space group with unique b axis, there is a possible origin shift of a half in x or z directions and any position along y) as is the choice of hand (x,y,z or -x,-y,-z). In the first instance, **VECSUM** calculates a symmetry function, calculating Harker vectors for every grid point in the asymmetric unit of the crystal, and looking up the value of the Patterson map. An initial site is chosen from the symmetry function, and subsequent sites were

found by evaluating the superposition function; again calculating Harker vectors, but also examining cross vectors between the first site and the current grid point being considered. The second site is fed in and the procedure repeated. After putative sites had been found, the refinement of these sites was then carried out using vector-space refinement as implemented in VECREF, and these refined sites input into MLPHARE [Otwinowski & Zbyszek, 1991]. MLPHARE performs further site refinement as well as calculating protein phases ϕ_P using the maximum likelihood method, which describes the phase ambiguity as a probability density function $P(\phi_P)$, with the maxima of this function lying at the most probable phase value. R_{cullis}, phasing power and Figure of Merit (FOM) statistics were monitored to assess the quality of the derivative. FFT was then used to calculate difference Fouriers using the reflection file from MLPHARE, and PEAKMAX [Evans, 1998] to search for additional sites which were in turn input into VECREF, and the iterative cycle repeated. An attempt to solve other derivatives (using the phases from the first) could then be carried out using the same methods.

5.4 Heavy Atom Data

All data (figure 5.4) were scaled to the monoclinic native data set (M1_F) but calculation of possible minor sites gave no reasonable phasing statistics (no phasing power above 1.0). The 7 day soak iodine (KI/I2) data; I3_4, xenon data; XE_2 and $K_2Pt(II)(NO_3)_4$ data; PT2_1, were all non-isomorphous with the native data (other derivative data sets were also non-isomorphous and could therefore not be used as a pseudo-native).

The $CdCl_2$ data; CD_6, K_2PtCl_4 data; PTCL4_4 and 24 hour iodine soak data; I3_3 were approximately isomorphous with the native but failed to have any major sites (displayed peaks no greater than 3 sigma on the Harker section). The iridium (Na_2IrCl_6) data IR_1 was promising as it was isomorphous with the native and had several peaks on the Harker section but none of these sites refined to produce good statistics. The Harker sections for these latter data sets are shown in figure 5.5.

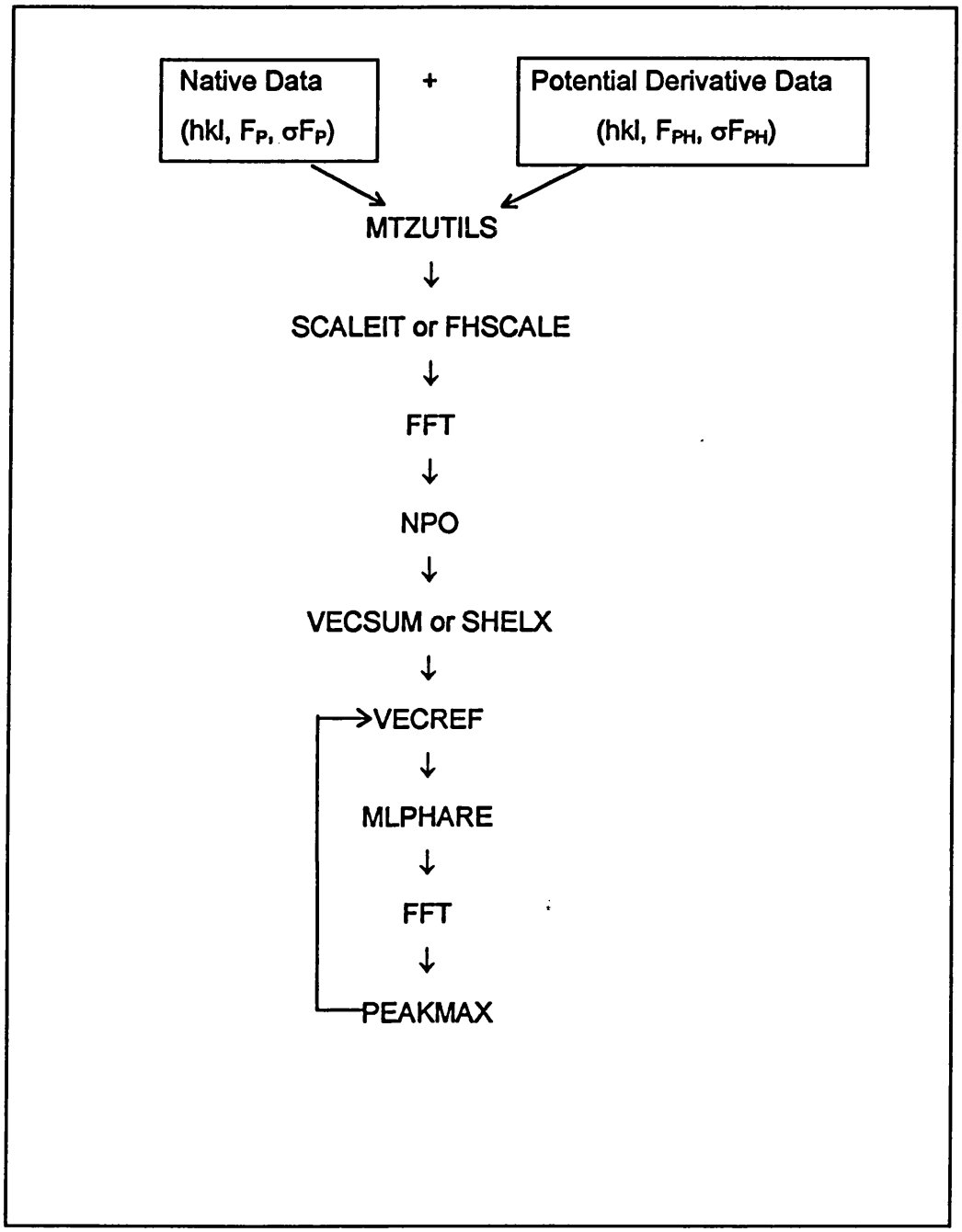
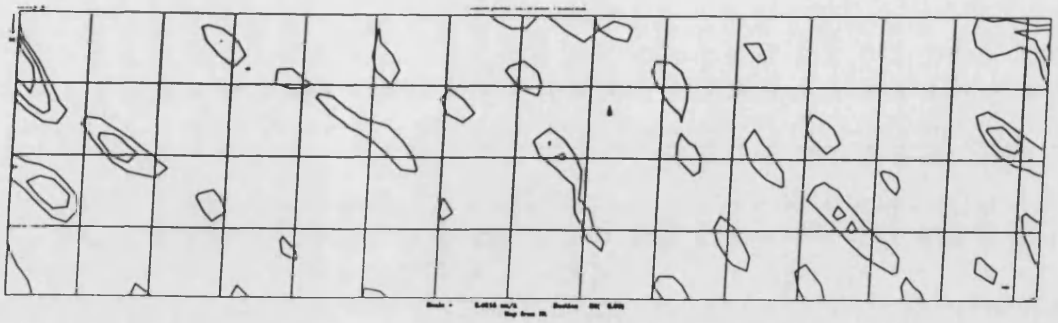


Figure 5.3: Scheme for analysis of heavy atom data.

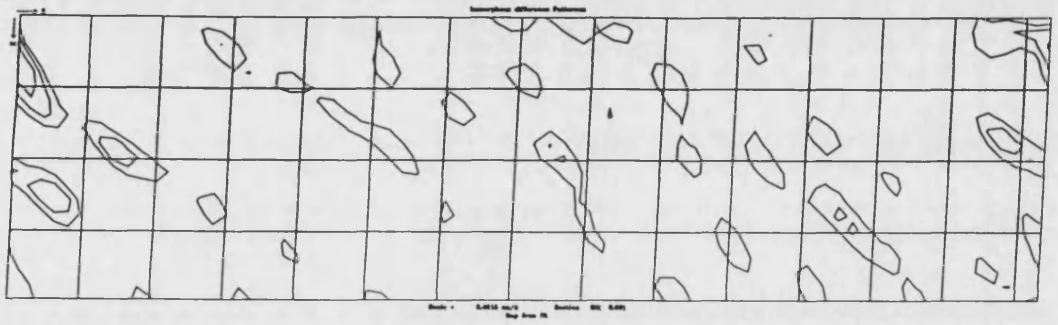
Data Set & Temperature of Collection	Space Group	Cell Dimensions /Å & °	Res. Limit /Å	Observed (unique) reflections	Completeness/ %	R_{merge} /%	I/sigmal	Top Shell /Å	R_I/%*
CD_6 (2CP, 100K)	$P2_1$	a=79.26, b=89.44 c=144.87, β=92.95	3.5	55347 (39017)	- (95.2)	- (18.8)	-	3.62-3.5	23.9
I3_3 (2CP, 100K)	$P2_1$	a=79.24, b=89.12 c=145.38, β=92.93	2.9	468754 (45117)	96.1 (94.0)	11.0 (22.6)	11.1 (5.3)	3.0-2.90	25.4
I3_4 (2CP, 100K)	$P2_1$	a= 79.76, b=88.66 c=144.86, β=92.38	4.2	97208 (14956)	99.7 (100)	8.3 (10.1)	15.0 (12.8)	4.35-4.20	not isomorphous
XE_2 (2CP, 100K)	$P2_1$	a=78.43, b=89.22 c=143.99, β=93.07	4.0	120699 (17235)	75.2 (78.2)	12.8 (16.8)	8.1 (7.7)	4.14-4.0	not isomorphous
IR_1 (2CP, 100K)	$P2_1$	a=78.99, b=89.02 c=145.36, β=92.8	6.5	38944 (4102)	99.7 (99.5)	13.3 (26.2)	11.9 (5.6)	6.71-6.5	22.6
PT2_1 (2CP, 100K)	$P2_1$	a=79.36, b=90.14 c=140.04, β=94.2	4.7	127297 (10549)	97.9 (94.5)	18.0 (26.1)	5.8 (3.1)	4.91-4.70	not isomorphous
PTCL4_4 (2CP, 100K)	$P2_1$	a=79.02, b=88.82 c=144.55, β=92.8	4.1	154289 (15888)	99.7 (99.0)	12.6 (16.1)	13.3 (10.9)	4.24-4.1	22.9
PTCL6_2 (2CP, 100K)	$P2_1$	a=79.03, b=88.82 c=144.78, β=92.90	3.5	199508 (25563)	93.9 (95.1)	9.9 (15.8)	13.2 (8.4)	3.62-3.5	28.7

Figure 5.4: Table summarising 'derivative' data sets collected (all data collected at 100K). Unless stated otherwise, values are defined for all reflections with those in parentheses corresponding to the top resolution shell (as defined in the final column). *R_I = Mean fractional isomorphous difference between a given derivative data set and the native set.

a)



b)



c)

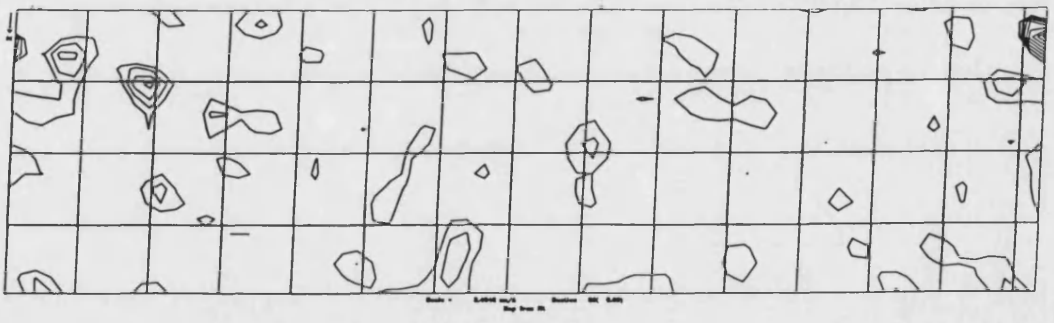
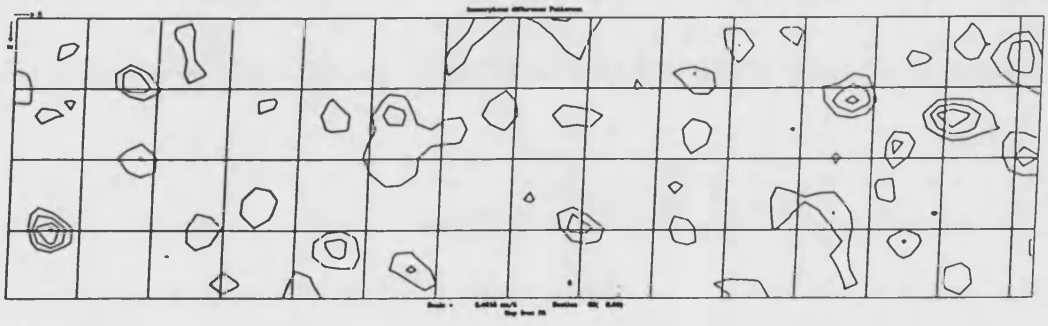


Figure 5.5: Harker sections (v = 1/2) for the 'derivative' data sets a) CD_6, b) I3_3, c) IR_1. a) and b) are calculated for data from 15-6Å and c) with data from 15-6.5Å.

The final heavy atom data set collected was the ($K_2Pt(IV)Cl_6$) 24 hour soak; PTCL6_2. It should be noted that the octahedral $[Pt(IV)Cl_6]^{2-}$ degrades in solution and tends to react as the square planar $[Pt(III)Cl_4]^-$. Crystals left to soak for two days in this reagent appeared to have cracked.

Data were scaled and the search for heavy atom sites was carried out as described in section 5.3. Scaling of the heavy atom to native data sets gave an R-factor of 19.1% in **SCALEIT**. **VECSUM** was used to search for heavy atom sites following calculation of a Difference Patterson with data from 15-6Å. The initial five sites were found using this method, and after refinement of these initial sites in **VECREF**, the remaining four sites were found by iterative calculation of difference Fouriers following rounds of **MLPHARE**. By this method, eight major sites and one weaker site were found (figure 5.7).

a)



b)

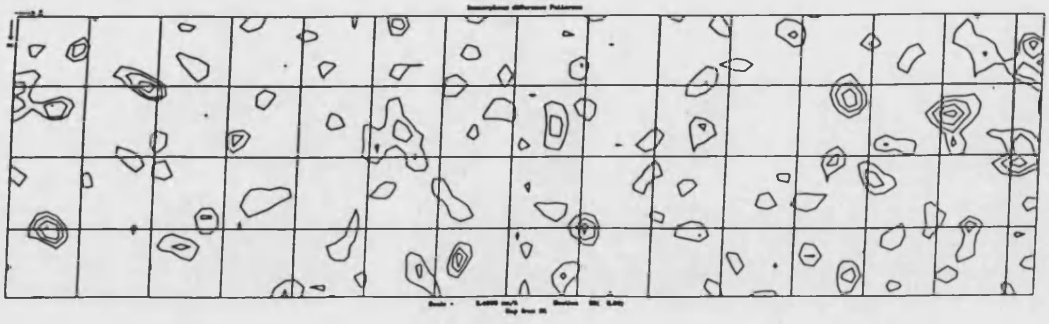


Figure 5.6: Harker sections (v=1/2) for the PTCL6_2 data calculated using data from a) 15-6Å and b) 15-3.5Å.

Site	x	y	z	Occupancy	B-factor/Å^2
1	4.19	34.29	69.94	0.622	20.0
2	42.83	38.19	143.06	0.565	20.0
3	39.99	56.84	137.97	0.622	20.0
4	6.60	15.74	70.95	0.596	20.0
5	3.16	20.78	58.55	0.619	20.0
6	11.73	29.30	75.24	0.557	20.0
7	47.19	52.03	147.39	0.602	20.0
8	39.18	42.56	131.61	0.611	20.0
(9)	61.47	21.12	69.10	0.400	20.0

Figure 5.7: Orthogonal coordinates of the nine $PtCl_4$ sites refined in MLPHARE.

After final refinement of the nine sites in MLPHARE the statistics suggested a reasonable derivative, with Phasing Power = 1.2, R_{cullis} = 0.64, and FOM = 0.32 (for all reflections) using data from 15 to 5Å. Attempts to calculate cross-phase difference Fouriers in attempt to solve the other derivatives proved unsuccessful and therefore this was an SIR case (Anomalous signal did not appear to be significant).

Calculation of an initial solvent flattened SIR electron density map from the phases calculated in MLPHARE (using the program DM [Cowtan, 1994]) revealed the presence of 'blobs' of density. Also, on analysis of all symmetry related Pt sites associated with these areas of higher density it could be seen that 4 Pt sites were associated with each tetramer in the asymmetric unit (sites 1,5,6,4 and 3,8,7,2 respectively) with one additional weak site (9). This was confirmed by the realisation that each of these sets above were related by the approximate translation known to relate one tetramer to the other (x+1/2, y+1/4, z+1/2). From the positions of these sites, the centre of mass of both the tetramers could be calculated using PDBSET [Evans, 1998] (assuming a single site per monomer). Knowledge of the NCS axes from the self rotation function allowed the NCS transformations to be defined (as both tetramers are in the same orientation) although there was ambiguity at this stage as to which

of the $P2_1$ related NCS sets were correct. The NCS symmetry transformations were calculated as below and are listed in figure 5.8.

As the transformation mapping monomer A onto monomer B is

$$X_B = R_B.X_A + T_B \qquad (1)$$

The translation vector, T_B can be calculated from knowledge of the rotation matrix, R_B (from the self rotation function). Monomer A is first shifted by the centre of mass, COM (using PDBSET) thus placing the tetramer at the origin,

$$X_B' = R_B.(X_A - COM)$$

$$\Rightarrow \quad X_B = X_B' + COM$$

Formula 1 can now be written as

$$R_B.(X_A - COM) + COM = R_B.X_A + T_B$$

$$\Rightarrow \quad T_B = COM - R_B.COM$$

This also holds for monomers C and D.

	ω	ϕ	κ	t_x	t_y	t_z
A1	64	180	180	55.18	50.06	113.99
A2	73	81	180	-0.419	-31.65	108.93
A3	32	318	180	-27.19	86.11	48.63
B1	64	180	180	126.59	94.8	259.54
B2	73	81	180	57.06	-79.06	226.20
B3	32	318	180	-8.80	178.90	78.89

Figure 5.8: NCS transformations for tetramers A and B.

A spherical mask (sphere radius 30Å) was created about the centre of both tetramers using the program NCSMASK which was trimmed for symmetry overlap (in this case input as $G2_1$ to account for both tetramers) and thus allowing 4-fold NCS averaging (within an approximate tetramer mask) in DM, treating each tetramer as a separate domain with phase extension from 5.0 to 2.7Å over 500 cycles. This resulted in a change in DM R_{free} from 66.4% to 34.9% and NCS correlation values from 0.25 to 0.85, (where statistics refer to before and after density modification respectively). A region of the averaged map can be seen in figure 5.9.

A main chain 'bones skeleton' was created from this DM averaged map in order to help identification of a single monomer, to allow a monomer mask to be created and eight-fold averaging to be used. The quality of the calculated averaged map was not high enough to get a satisfactory superposition of the model onto the map and therefore the approximate centre of the monomer was simply estimated using O, and a sphere of radius 15Å was created about the centre with NCSMASK (and again trimmed for both NCS and crystallographic symmetry). This was used for an initial mask to carry out eight-fold averaging in DM. In addition, the radius of the averaging mask was altered, and attempts were made to superimpose a TIM dimer model onto the density obtained after averaging, but none of the above methods produced better quality electron density maps.

5.5 Discussion

The methods of heavy atom co-crystallisation and soaking have been employed in an attempt to use Isomorphous Replacement to determine the *Pw*TIM structure. This derivative search was hindered on several counts; the fragile nature of the *Pw*TIM crystals meant that they were very susceptible to damage simply from handling, addition of heavy atom reagents, or the subsequent cryoprotecting and freezing process. Difficulties reproducing diffraction quality crystals in high enough quantity also slowed the screening process (especially with such a high crystal mortality rate). Although some useful phase information has now been obtained from the single $PtCl_6$

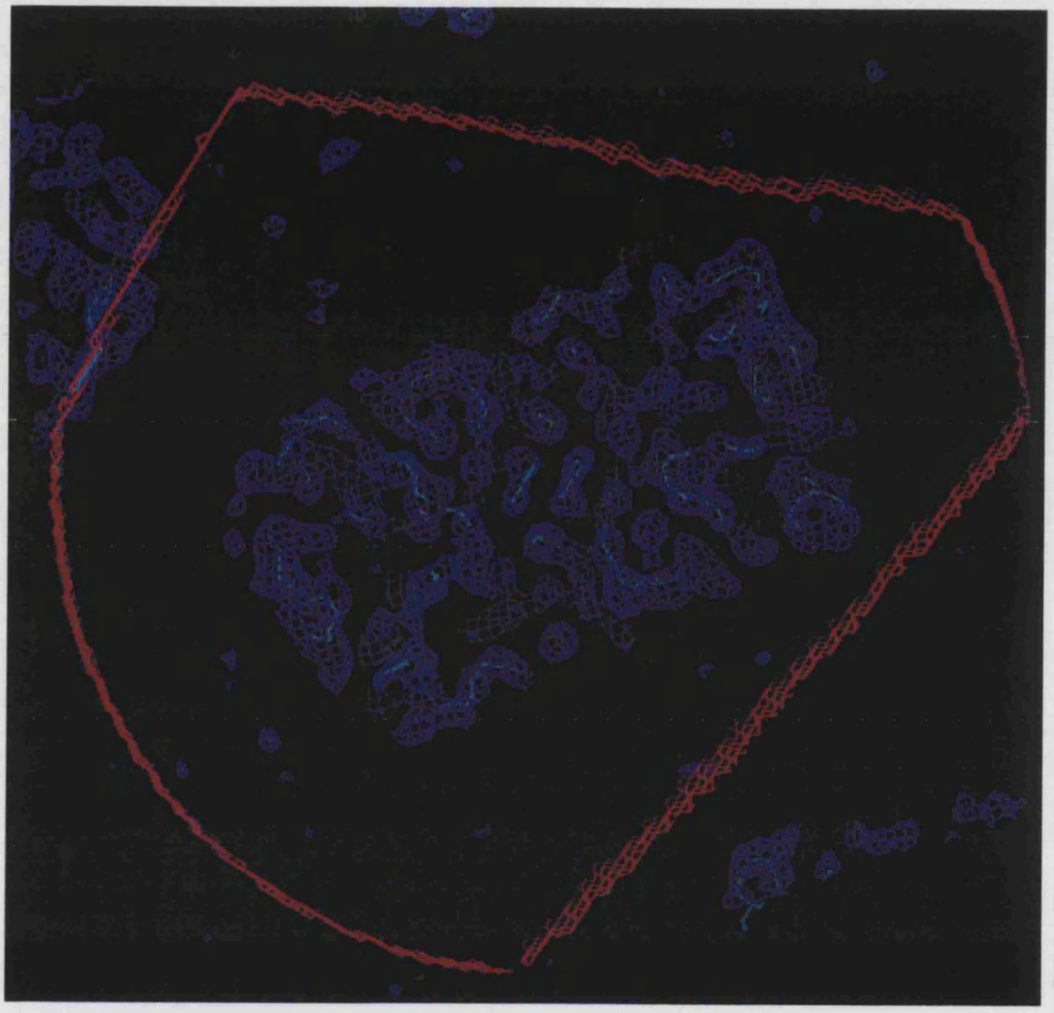


Figure 5.9: Section of the averaged electron density map contoured at 1σ (blue) and superimposed onto a bones skeleton shown inside the 30Å 'tetramer' mask (red). Displayed in **O**.

derivative, it looks likely that further experimentation will be needed to obtain another good derivative.

It was hoped that eight-fold NCS symmetry averaging would enable us to break the phase ambiguity of this SIR derivative. Although the centre of mass has been established for the two tetramers in the asymmetric unit (assuming identical substitution site on each monomer of the tetramer), the quality of the SIR/solvent flattened electron density map was not sufficient to enable a monomer mask to be created (which would have facilitated true eight-fold

averaging). Application of the NCS operators to carry out four-fold averaging separately within the 'tetramer' masks did not clarify the situation. Other methods such as the use of Cross-Crystal Averaging as implemented in DMMULTI [Cowtan, 1994], with the second monoclinic data set; M2_F (which extends to 3.0Å) may yet yield better quality maps, but time limits have prevented further studies on this line of work.

Different reagents (particularly species with lower activities), reagent concentrations, optimising length of soak times and pHs are all variables which may allow control of substitution rates and thus prevent crystal cracking. In addition to the follow up work required to obtain a suitable derivative, an attempt to obtain crystals of the SeMetTIM may also prove fruitful.

CHAPTER 6

Citrate Synthase from *Sulfolobus solfataricus*

6.1 Crystallisation

Crystallisation trials were set up (by Rupert Russell) using the hanging drop vapour diffusion method (as described in chapter 2) with the native *Ss*CS (in the presence of citrate and CoA) which had been overexpressed in *E. coli* [Connaris *et al*, 1998] and purified with a red gel affinity column as described by Muir *et al* [Muir *et al*, 1995]. The crystal yielding condition was a 6µl hanging drop containing 2µl of protein at a concentration of 10 mg/ml (with mM concentrations of the substrates), 2µl of TRIS/HCl pH 7.2 containing 17% PEG 8K and 2µl of 0.1M $CaCl_2$. A single rod like crystal of approximate dimensions 2mm x 0.1mm x 0.1mm was found in a partially dried out drop after a six month period. This was the only crystal obtained which was used for data collection purposes.

6.2 Crystallographic Data

No successful freezing conditions had been found previously and as only one crystal was available, data were collected at room temperature. The crystal was mounted in a quartz capillary tube and data collected on a 30cm Mar image plate detector (as in chapter 3). The crystal to detector distance was set to 220mm and contiguous frames recorded, rotating the crystal 0.5° ($\Delta\phi$) for ten minute exposures. Data extended to 2.6Å although were very weak at higher resolution and therefore only processed to 2.7Å (statistics shown in the table in figure 6.1). As the crystal gradually suffered from radiation damage it was translated stepwise, perpendicular to the beam to maximise completeness of the data set. The data were reduced as described earlier using the DENZO/SCALEPACK software. Initial indexing suggested a primitive monoclinic lattice but no (0k0) reflections were recorded to distinguish P2 and $P2_1$ space groups. Molecular Replacement was carried out in both space

groups (although it was assumed that it was more likely to be $P2_1$ due to the former being rarely observed for proteins) and was later confirmed to be $P2_1$ on refinement of the structure.

Space Group	$P2_1$
Unit Cell Dimensions	a=77.34Å, b=97.86Å, c=119.33Å, β=107.6°
Resolution limit	2.7Å
Data Completeness	88.6% (91.2%)
R_{merge}	7.2% (22.9%)
I/sigmal	9.37 (3.25)
Total No. of reflections	148169
Unique Reflections	46758

Figure 6.1: Table displaying native data collected for the *Ss*CS (disk file name 21oct_6). Data in parentheses correspond to the high resolution data shell (2.82-2.71Å)

The solvent content (V_{solv}) was calculated using the Mathews approximation (chapter 3) and suggested the likelihood of there being two dimers in the asymmetric unit (V_{solv}=51%) as only one dimer in the asymmetric unit gave a relatively high solvent content (V_{solv}=74%).

6.3 Molecular Replacement

As with the *Pw*TIM, there were several known structures available as phasing models for Molecular Replacement (the coordinate files are listed in figure 7.1) of which the *T. acidophilum* sequence had the highest homology (59% identical). It was assumed that the *Ss*CS had been crystallised in the 'closed' form, with the crystallisation being in the presence of substrates (the small domain is known to rotate relative to the large domain on binding of substrates). Initial Molecular Replacement was therefore carried out with monomers and dimers of the *Pf*CS and *DS*CS structures as search models (as the *Ta*CS was only available in the 'open' form). Cross rotation and translation

searches were carried out in the program AMoRe (the process being described in more detail in chapter 4) with the translation function carried out for P2 and $P2_1$ space groups. No obvious solutions were found with these search models. The *Ta*CS dimer was then used as a model with data from 15-6Å included in the rotation function and an integration radius of 2.5-25Å. The translation function was calculated for the 50 top solutions from the rotation function using data from 15-6Å and the top solution (number 33 from the rotation search) which had a correlation coefficient of 32.0 and R-factor of 53.4% (compared with a background correlation coefficient of 23 and R-factor of 56%). This solution was fixed and the translation search run again for the remaining rotation function solutions using data in the same resolution range. A solution for the second dimer (solution number 10 from the rotation search) had a correlation coefficient of 37.7 and R-factor of 51.9% (background correlation coefficient 31 and R-factor of 53). A rigid body refinement was carried out in AMoRe (simply treating the dimers as rigid bodies) and the final solutions had a correlation coefficient of 56.6 and R-factor of 41.3% (background correlation coefficient of 27 and R-factor of 52). This easily obtained solution found using the *Ta*CS open structure as a phasing model (when attempts to use the closed structures had failed) roused suspicion as to the presence of substrates in the active site of the enzyme.

Solution	α	β	γ	tx	ty	tz
1	184.6	46.4	158.6	0.114	0.0	0.469
2	224.8	62.1	120.2	0.832	0.439	0.943

Figure 6.2: Molecular Replacement solutions for the two dimers after rigid body refinement in AMoRe (defined in Eulerian angles and fractional coordinates).

6.4 Refinement and Model Building

6.4.1 Introduction

The restrained Refinement of the *Sulfolobus solfataricus* Citrate Synthase structure was carried out using the program **REFMAC** [Murshudov *et al*, 1997], minimising the Maximum Likelihood Residual by the Conjugate Gradient Sparse Matrix method [Tronrud, 1994]. Noise filtering phase improvement techniques (NCS averaging [Bricogne, 1974], solvent flattening [Wang 1985, Leslie 1988] and histogram matching [Zhang & Main, 1990]) as implemented in the program **DM** were used to improve the quality of electron density maps. Stereochemical restraints were imposed using **PROTIN** [Hendrickson, 1985]. The initial R-factor in **REFMAC** (after rigid body refinement) was 48.3% (R_{free}=48.6%) and final R-factor of 20.8% (R_{free}=28.5%) for all data from 20.0Å to 2.7Å. With data to only 2.7Å, the resolution was not sufficient to include structural waters in the final model. In addition, the low ratio of observations to parameters meant that the isotropic refinement of B-factors had to be approached with some caution. This was however carried out in the later stages of refinement by applying tight four-fold NCS restraints. The process is described in more detail below and summarised in figure 6.3.

6.4.2 Refinement Method

Rigid body refinement;

Rigid body refinement was carried out in **AMoRe** (treating each dimer as a rigid body) after the initial Molecular Replacement solutions for the two dimers in the asymmetric unit had been found.

Calculation of structure factors;

After rigid body refinement, initial structure factors were calculated using the experimentally observed amplitudes and phases from the correctly oriented *Ta*CS models (polyalanine except where there were conserved side-chains in the *Ss*CS sequence) using **SFALL** [Dodson & Baker, 1998]. Prior to this step, missing number flags were assigned in **MTZMNF** [Winn & Dodson, 1998] to

account for the incomplete data followed by generation of a complete list of indices in UNIQUE [Leslie, 1998], with FREERFLAG flagging 5% of the data for use as a test set (FreeRflags were merged with the data using CAD [Dodson, 1998]). This test set was used to calculate an R_{free} [Brunger, 1992b] value which is not affected by the refinement procedure (as they are not used in the refinement). SIGMAA [Read, 1986] was then used to establish weights for the structure factors calculated in SFALL based on a comparison of the experimentally observed magnitudes with those calculated for the model.

Density modification;

This reflection file output from SIGMAA containing weighted map coefficients and FOMs was input into DM for 10 cycles of density modification incorporating four-fold NCS averaging, solvent flattening (based on the known solvent content as described in section 6.2) and histogram matching. The protein-solvent boundary was defined by a mask created using a 5Å radius around the initial *Ta*CS model in NCSMASK. Conversion matrices mapping monomers A onto B,C and D were supplied using LSQKAB. Success of the density modification procedure was monitored by the DM R_{free} and correlation coefficients of the NCS matrices. During the initial 10 cycles the DM R_{free} value dropped from 44% to 30.4% and NCS matrices refined to give NCS correlation values of approximately 0.8 (from 0.6 at the beginning of the density modification procedure.

Initial calculation of density modified map;

FFT was used to calculate an electron density map from the DM phases which was extended over a region covering one monomer in EXTEND and subsequently reformatted in MAPMAN [Kleywegt & Jones, 1996] to be read into O in order to carry out manual building. Rotamer conformations for those side-chains built into a single monomer of the *Ta*CS model were checked and side-chains with no density in this map were removed (contoured above 2σ). The initial maps obviously contained an element of bias due to the inaccuracies of the phases for this model (and also from building into subsequent maps) but an attempt to compensate for this was made by the weighting on Fo and Fc as applied in SIGMAA. The monomer coordinates

were read out from O and NCS matrices used to create the other 3 monomers (using **LSQKAB** and **PDBSET**).

Restrained positional refinement;

Geometric restraints could then be set up for the model (ABCD) using the program **PROTIN** based on a dictionary of ideal protein stereochemistry as observed in high resolution small molecule structures [Priestle, 1995, Engh & Huber, 1991]. This creates a restraints file corresponding to the model which is input into **REFMAC** for positional refinement. Tight NCS restraints for both main-chain and side-chain were used initially and 6 cycles of refinement carried out in **REFMAC** with a weight of 0.5 using the average diagonal term of X-ray and Geometry with experimental sigma values included in the weighting. Positional refinement was carried out with an initial B-factor of 20Å^2 (single overall B-factor refined for whole model) and a bulk solvent correction [Tronrud, 1997] included. After the first round of refinement the R-factor was 36.3% (R_{free}=40.5%).

Iterative manual model building and refinement;

The new (refined) model and reflection file (containing **SIGMAA** style weighted coefficients) output from **REFMAC** were used to calculate weighted mFo-DFc and 2mFo-DFc difference maps (as before using **FFT** and **EXTEND**) for subsequent manual building steps in O (again the maps were contoured above 2σ). In addition, the coordinates for the refined model along with an updated solvent mask and NCS matrices were put back into **SFALL** to calculate new Structure Factors for the following phase improvement step and the iterative process then continued. After the initial round of refinement, the three deletions with respect to the *Ta*CS sequence could also be removed (corresponding to the four residues from three loop regions; Q51, N116 and R332, N333 in the *Ta*CS sequence) and the main-chain remodeled in these regions. Side-chains were then built into a single monomer (choosing the appropriate rotamer conformations) of the model in O mainly using the DM and 2mFo-DFc maps over successive cycles to an R-factor of 28.9% (R_{free} 33.8%). Significant remodeling of the main-chain backbone was not required (except for several surface loop regions) due to the high homology of the starting model.

B-factor refinement;

Keeping the tight NCS restraints, individual isotropic B-factor refinement was then carried out, bringing the R-factor down to 24.7% (R_{free} 31.2%) after which the NCS restraints were gradually loosened and the four monomers were built independently. NCS restraints were controlled in **PROTIN** and during the refinement procedure side-chain followed by main-chain restraints were gradually loosened, with a final round removing the NCS restraints continuing to lower the R_{free} value.

Model analysis and further building;

Areas of bad geometry identified by **PROCHECK** [Laskowski *et al*, 1993] and the Real Space Refinement program **OOPS** [Kleywegt & Jones, 1996b] were scrutinised more closely in the latter stages of refinement. These programs highlight regions of poor stereochemistry and **OOPS** also identifies areas of low correlation between map and model. Weighting on the geometry was then tightened by reducing the factor on the diagonal terms from 0.5 to 0.2. Two extra residues at the N terminus and five extra residues at the C terminus (compared with *Ta*CS model) were also built into the final model (with the final model comprising the four monomer chains A3-A369, B3-B368, C3-C370 and D3-D367). The first two residues at the N-terminus and last seven residues of the C-terminal arm (discussed later) were not seen in the poorly defined electron density of these parts of the structure. One conflict with the sequence data was residue 57 which had been assigned as arginine and was found from analysis of the electron density map to be a proline (this is a totally conserved proline in all the other known citrate synthases). The superposition of the active site residues of the small domain of *Ss*CS and *Ta*CS and the absence of density for substrates (in mFo-DFc and 2mFo-DFc maps) in the active site support the previous speculation that the structure is the 'open' form of the enzyme. After 24 rounds of refinement in **REFMAC** the final R-factor was 20.8% (R_{free}=28.5%).

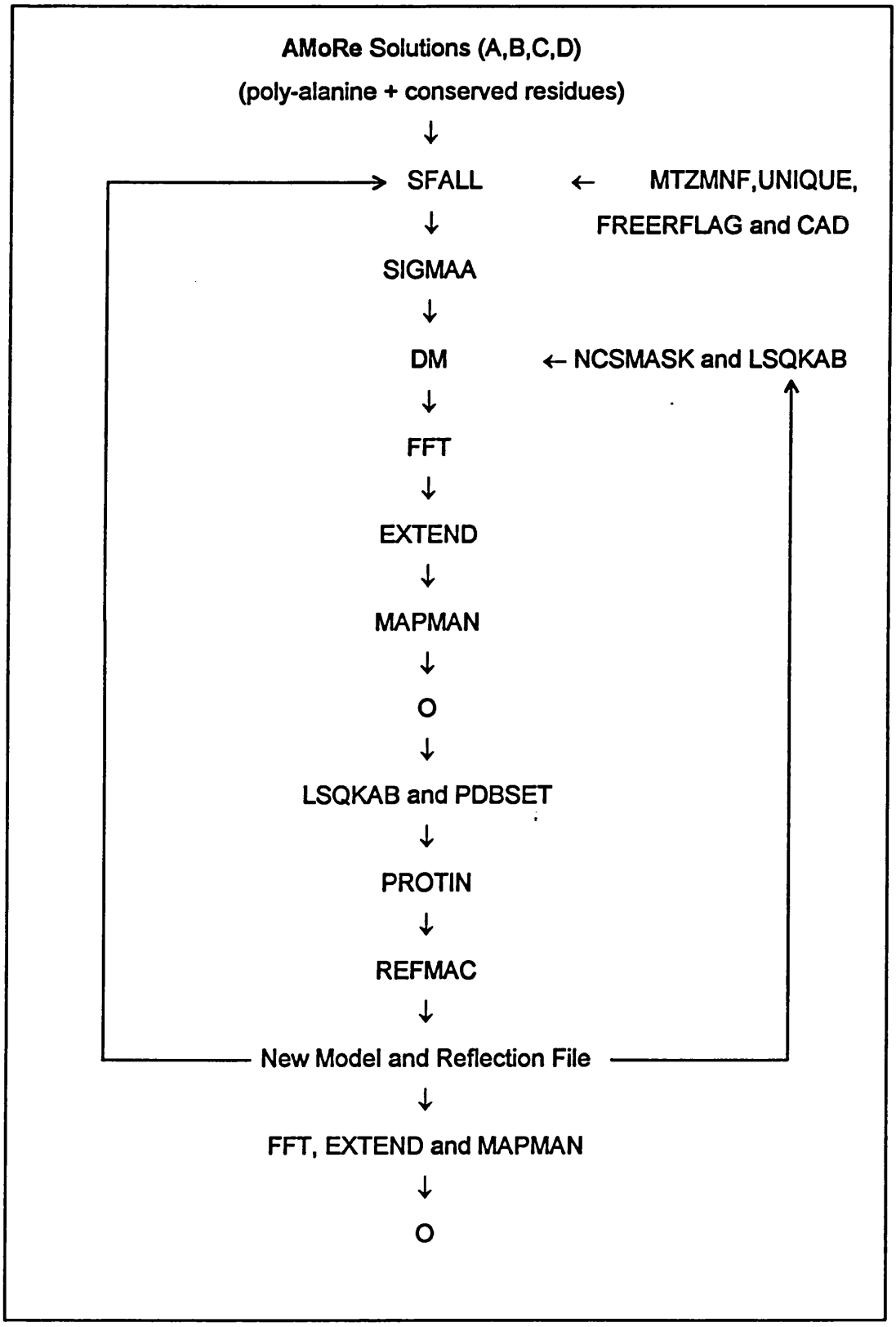


Figure 6.3: Scheme summarising the refinement procedure

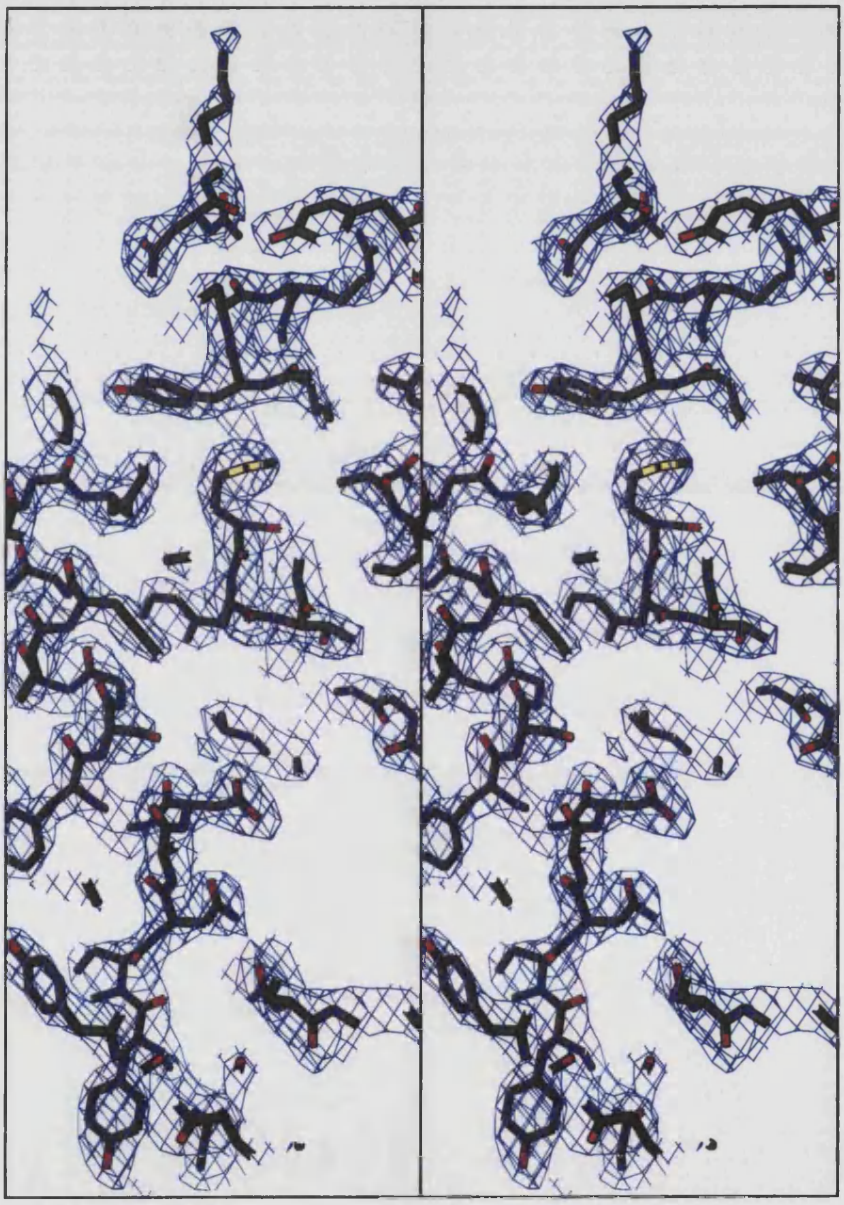


Figure 6.4: Stereo-diagram showing a typical region of the final 2Fo-Fc electron density map contoured at 1σ (diagram created using **BOBSCRIPT**).

Observations	148169
Unique reflections	46758
R-Factor /%	20.8
Free R-Factor /%	28.5
No. of protein atoms	11742
No. ligand atoms	0
No. solvent atoms	0
RMSD bond lengths /Å	0.009
RMSD bond angles (defined as distance) /Å	0.032
RMSD planar 1-4 distance /Å	0.032
Overall G-factor (PROCHECK)	-0.06
Overall coordinate ESU	0.444

Figure 6.5: Refinement statistics

6.5 Validation of Structure

Although weighted stereochemical restraints were applied during the refinement procedure, stereochemistry of the final model was analysed using the program PROCHECK. Figure 6.9 contains various outputs from PROCHECK for the four monomers in the asymmetric unit and some other refinement statistics are listed in figure 6.5. The general overview of main-chain and side-chain parameters including bonded and non-bonded interactions are better than or inside deviations from ideal geometry expected for the resolution to which the structure was refined (including the overall G-factor of -0.06). Plots showing deviation of main-chain parameters from that of small molecules show several features; virtually all (99.9%) of main-chain bond lengths (only 2 N-terminal residues differ by more than 0.05Å) are inside the 2σ RMS deviation limits from the ideal values. Main-chain bond angles are also well restrained with 87.1% lying inside the limits. Planarities listed for ring and non-ring planar groups show the general low RMS distances from planarity (only 0.2% outliers). Plots of side-chain torsion angles show that $\chi1\chi2$ distributions are clustered about the mean values with no serious steric

clashes. The Ramachandran plot shows the main-chain phi-psi torsion angles for the four monomers in the asymmteric unit. 91.3% of residues lie in the most favoured regions, with only 8.7% in the allowed regions and no residues appearing in the generously allowed or disallowed regions (excluding glycine and proline residues). Glycine and Proline residue Ramachandran plots show that there are only a few residues slightly outside accepted torsion angles.

Included in the residue property outputs are the B-factor plots (plotted per residue) which show the areas of disordered (or misbuilt) structure. A summary of the B-factors for the four monomers is diplayed in figure 6.6. Comparison of the temperature factors for all four monomers show that they all have corresponding regions of highest B-factors (and this is similar to the trend found for other CSs compared in chapter 7). The regions of highest B-factor comprise mainly of surface or active site loops; notably residues 57-63 (loop D and the first part of helix E), 107-112 (loop G-I), 162-166 (loop J-K), 245-253 (first part of loop O-P), 279-282 (loop P-Q), 302-308 (centre of loop Q-R) and the end of the C-terminal arm. These are regions of the structure which one would expect to see the greatest flexibility, especially as this is the open form of the enzyme. Despite this common trend between the four monomers, there is a considerable difference in values of the average B-factors. Monomers A and C have a comparable and significantly higher average B-factor than those of monomers B and D (which are also comparable with each other). The deviation is greatest in the small domain and as this is the most flexible region of the structure. In particular, B-factors for the small domain of monomer B are considerably lower than that of the other three monomers. Examination of the crystal packing for the four monomers may explain this feature. Numbers of crystal contacts for crystallographic and non-crystallogrphic symmetry related monomers were calculated using a distance from 0-3.6Å (using the program **CONTACT** [Skrzynski, 1998]). These calculations show that the number of residues involved in crystal contacts are significantly increased only for monomer B (A, 12 residues, B, 23 residues , C, 11 residues and D has13 residues). Analysis of the different overall environments within the crystal as shown in figure 6.10 also suggests that monomers A (yellow) and C (cyan) do appear to be less closely associated with the neighbouring protein, and are surrounded by large solvent channels.

	A	B	C	D
Average B/Å^2	42.8	31.0	41.1	34.2
Minimum B/Å^2	18.5	12.0	15.6	13.6
Maximum B/Å^2	86.0	67.6	87.6	67.9
Average B /Å^2 (large domain)	38.1	30.8	37.8	30.3
Average B/Å^2 (small domain)	55.3	34.8	49.4	43.6

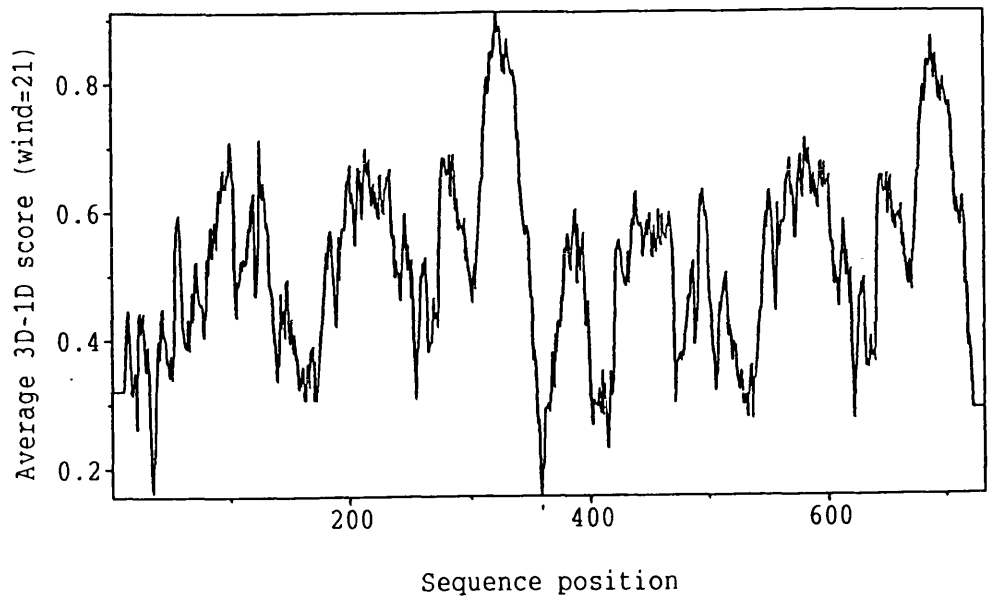
Figure 6.6: B-factors for the four *Ss*CS monomers

The program ENVIRONMENTS [Luthy *et al*, 1992] was run to analyse the chemical sense of the model. This involves calculating a matrix of scores based on the burial or exposure of given side-chains with relation to their chemical nature (over a window size of 21 residues). 3D-1D profile plots (figure 6.8) from these matrices was produced for both dimers and shows clearly that there are no regions of the structure in which residues are in totally unacceptable chemical environments (all scores above zero).

MONOMER	A	B	C	D
A	—	0.847	0.953	0.865
B	—	—	0.855	0.784
C	—	—	—	0.916

Figure 6.7: LSQ fits for alpha-carbon atoms of the four *Ss*CS monomers calculated using LSQKAB and given in Angstroms.

a)



b)

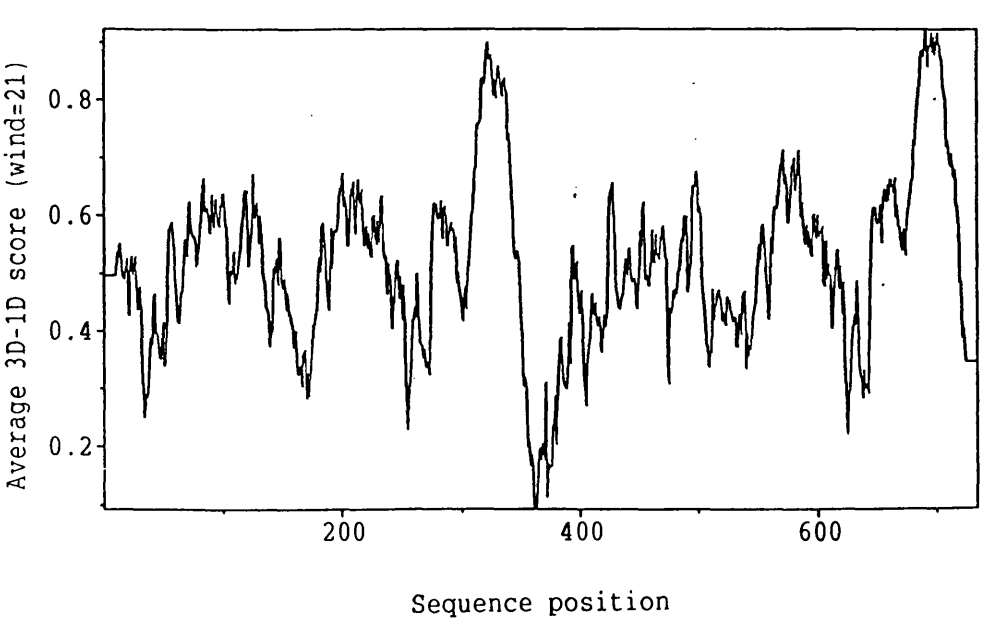
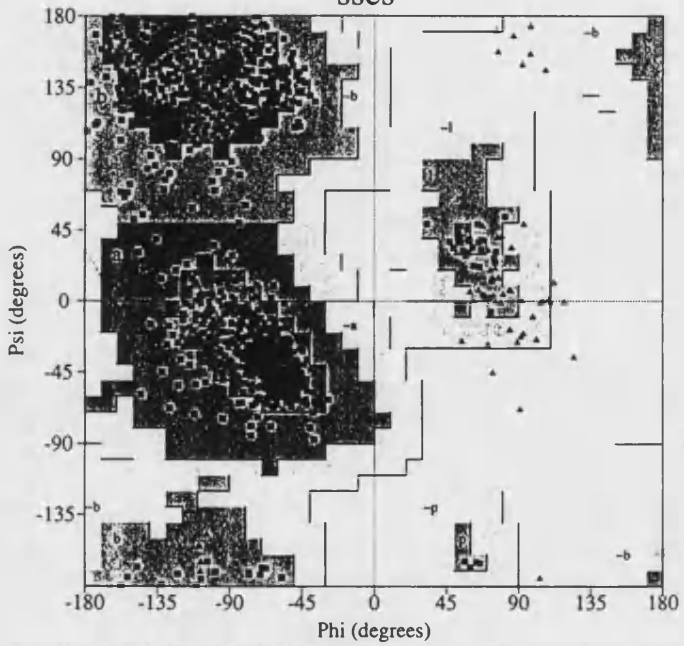


Figure 6.8: 3D-1D profile plot for dimers a) AB and b) CD created with ENVIRONMENTS.

PROCHECK

Ramachandran Plot

sscs



Plot statistics

Residues in most favoured regions [A,B,L]	1201	91.3%
Residues in additional allowed regions [a,b,l,p]	115	8.7%
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0%
Residues in disallowed regions	0	0.0%
Number of non-glycine and non-proline residues	1316	100.0%
Number of end-residues (excl. Gly and Pro)	8	
Number of glycine residues (shown as triangles)	84	
Number of proline residues	60	
Total number of residues	1468	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

sscs_01.ps

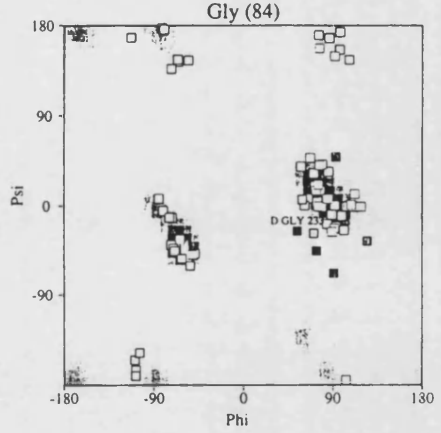
PROCHECK

Page 1

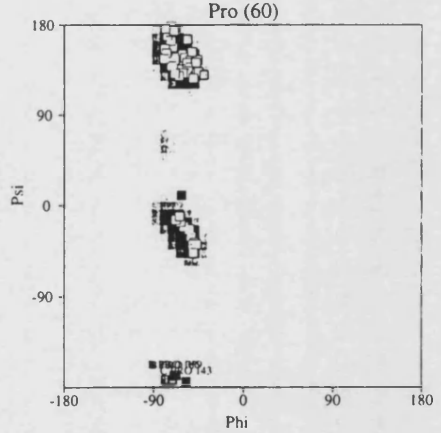
Ramachandran plots for Gly & Pro

sscs

Gly (84)



Pro (60)



Numbers of residues are shown in brackets. Those in unfavourable conformations (score < -3.00) are labelled.
Shading shows favourable conformations as obtained from an analysis of 163 structures at resolution 2.0A or better.

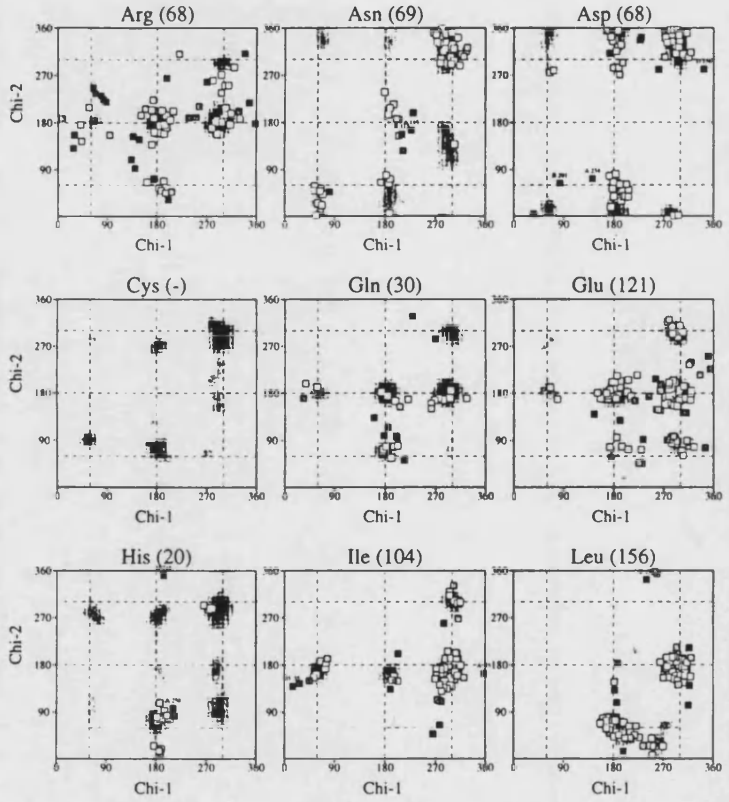
sscs_02.ps

PROCHECK

Page 1

Chi1-Chi2 plots

sscs



Numbers of residues are shown in brackets. Those in unfavourable conformations (score < -3.00) are labelled.
Shading shows favourable conformations as obtained from an analysis of 163 structures at resolution 2.0A or better.

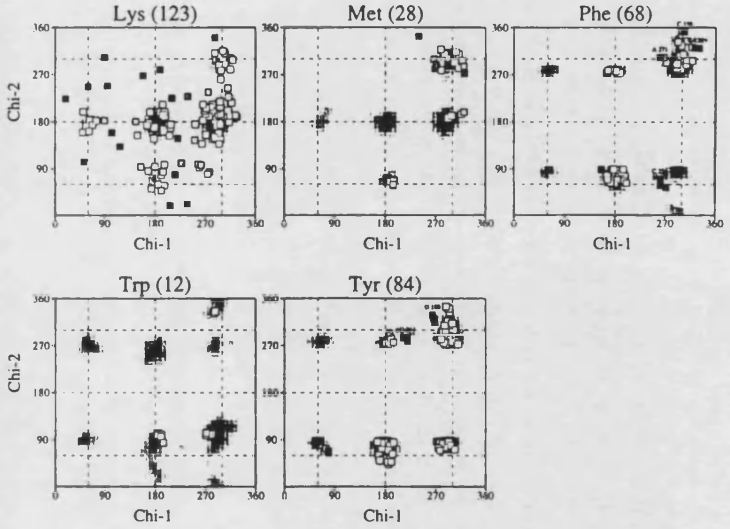
sscs_03.ps

PROCHECK

Page 2

Chi1-Chi2 plots

sscs



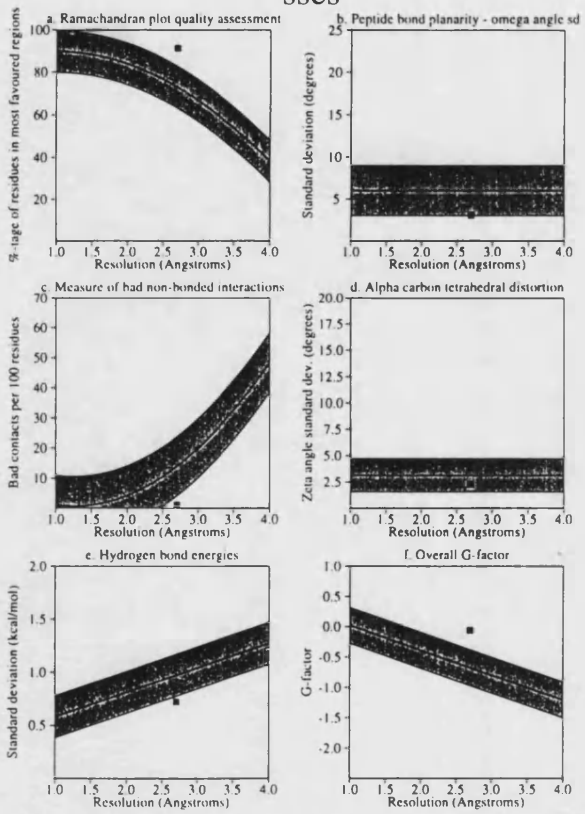
Numbers of residues are shown in brackets. Those in unfavourable conformations (score < -3.00) are labelled.
Shading shows favourable conformations as obtained from an analysis of 163 structures at resolution 2.0A or better.

sscs_03.ps

PROCHECK

Main-chain parameters

sscs



Plot statistics

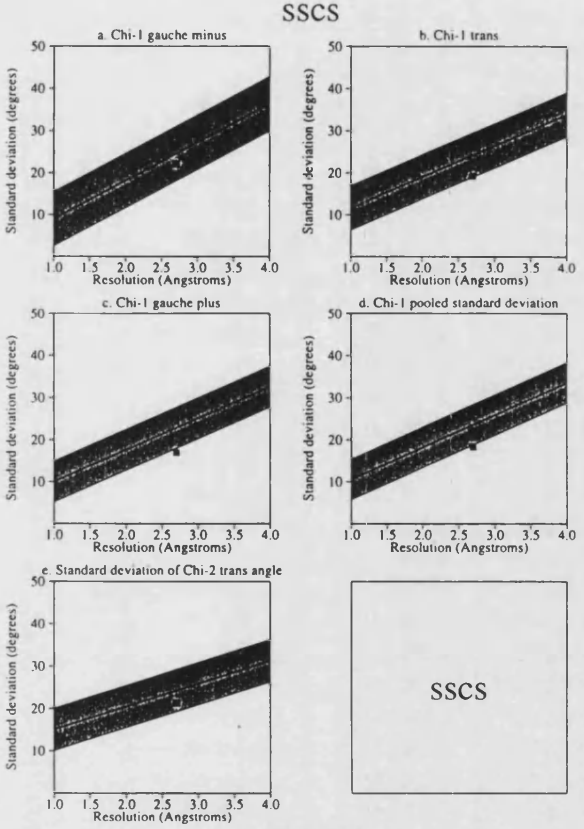
Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. %-tage residues in A, B, L	1316	91.3	72.9	10.0	1.8	BETTER
b. Omega angle st dev	1464	3.1	6.0	3.0	-1.0	Inside
c. Bad contacts / 100 residues	17	1.2	13.9	10.0	-1.3	BETTER
d. Zeta angle st dev	1384	2.3	3.1	1.6	-0.5	Inside
e. H-bond energy st dev	967	0.7	1.0	0.2	-1.2	BETTER
f. Overall G-factor	1468	-0.1	-0.7	0.3	2.0	BETTER

sscs_04.ps

PROCHECK

Side-chain parameters

sscs



Plot statistics

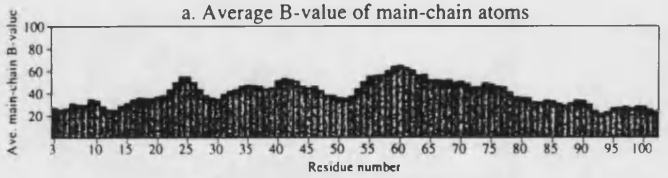
Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. Chi-1 gauche minus st dev	115	22.1	24.5	6.5	-0.4	Inside
b. Chi-1 trans st dev	407	19.1	24.2	5.3	-1.0	Inside
c. Chi-1 gauche plus st dev	654	16.9	22.8	4.9	-1.2	BETTER
d. Chi-1 pooled st dev	1176	18.3	23.5	4.8	-1.1	BETTER
e. Chi-2 trans st dev	423	21.2	24.2	5.0	-0.6	Inside

sscs_05.ps

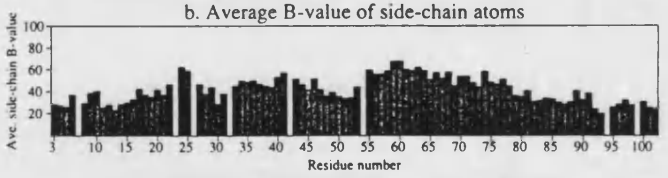
Residue properties

sscs

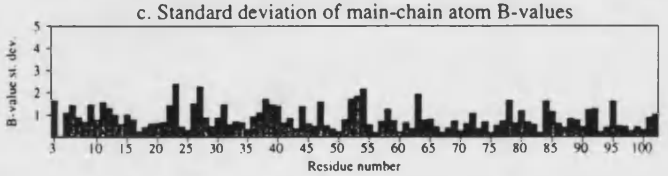
a. Average B-value of main-chain atoms



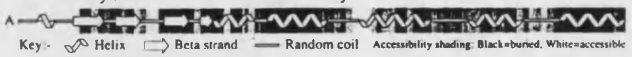
b. Average B-value of side-chain atoms



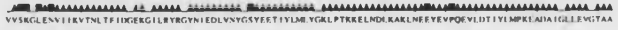
c. Standard deviation of main-chain atom B-values



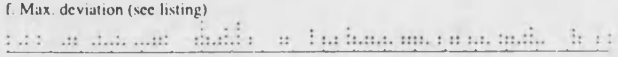
d. Secondary structure & estimated accessibility



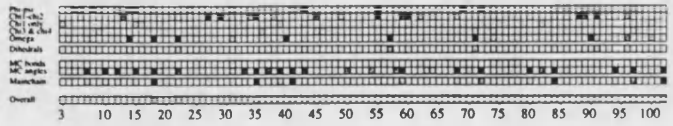
e. Sequence & Ramachandran regions ▲ Most favoured ■ Allowed ▮ Generous ▮ Disallowed



f. Max. deviation (see listing)



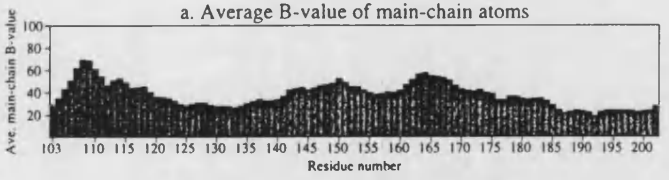
g. G-factors



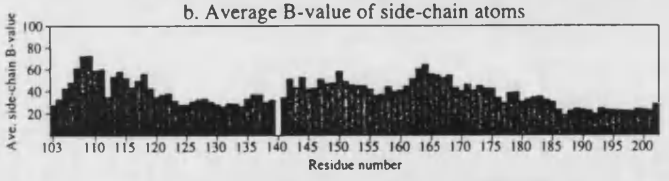
Residue properties

sscs

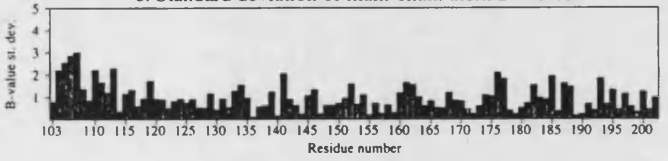
a. Average B-value of main-chain atoms



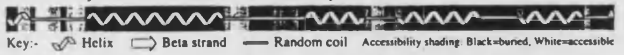
b. Average B-value of side-chain atoms



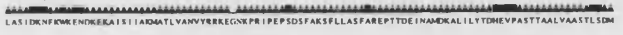
c. Standard deviation of main-chain atom B-values



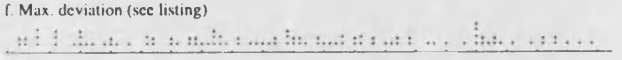
d. Secondary structure & estimated accessibility



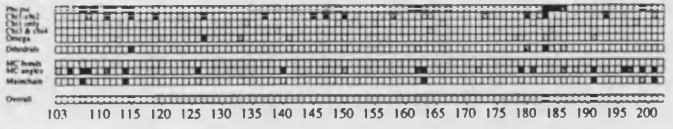
e. Sequence & Ramachandran regions ▲ Most favoured ■ Allowed ▮ Generous ▮ Disallowed

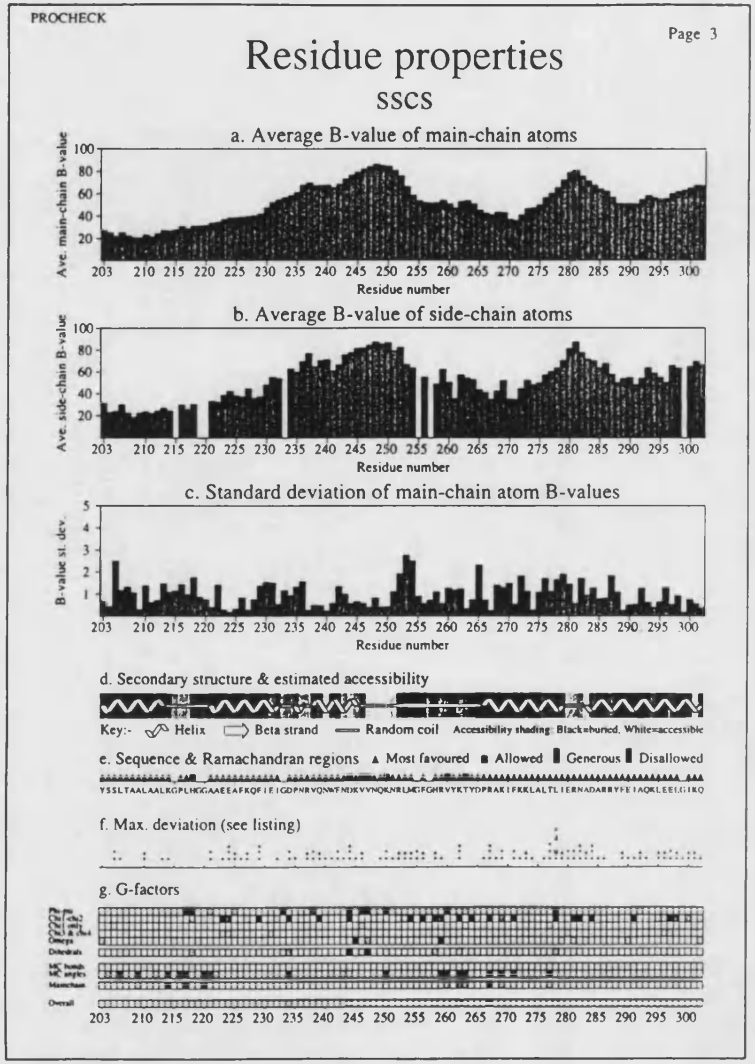


f. Max. deviation (see listing)

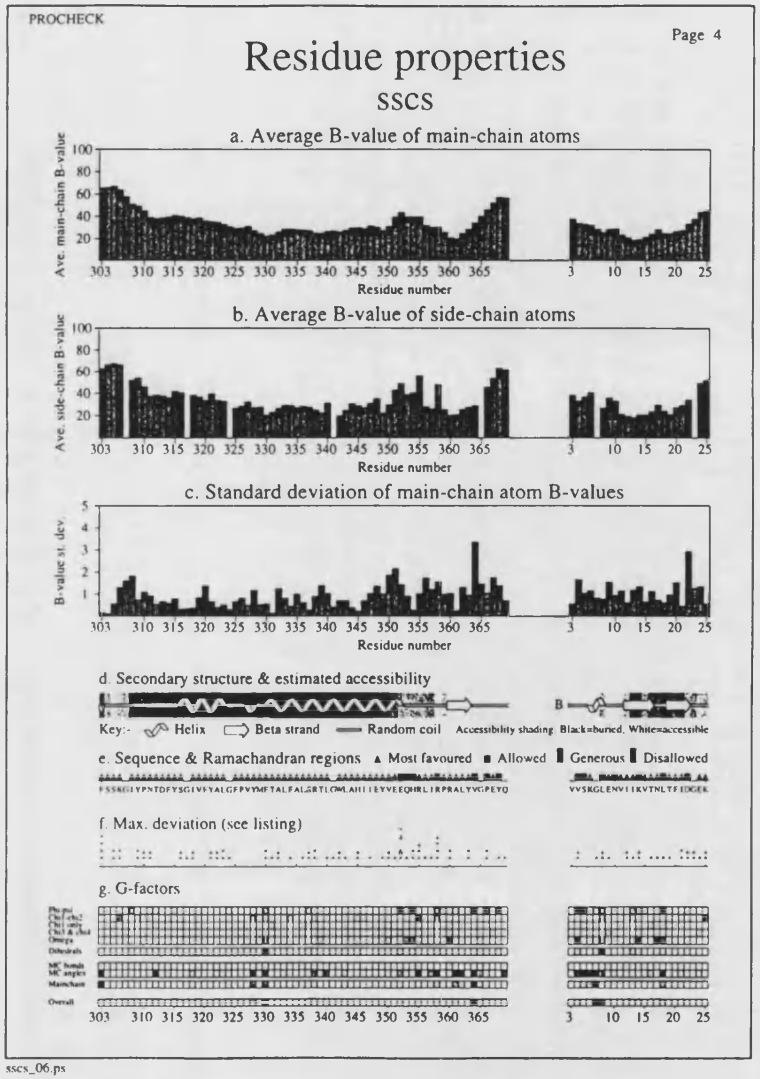


g. G-factors

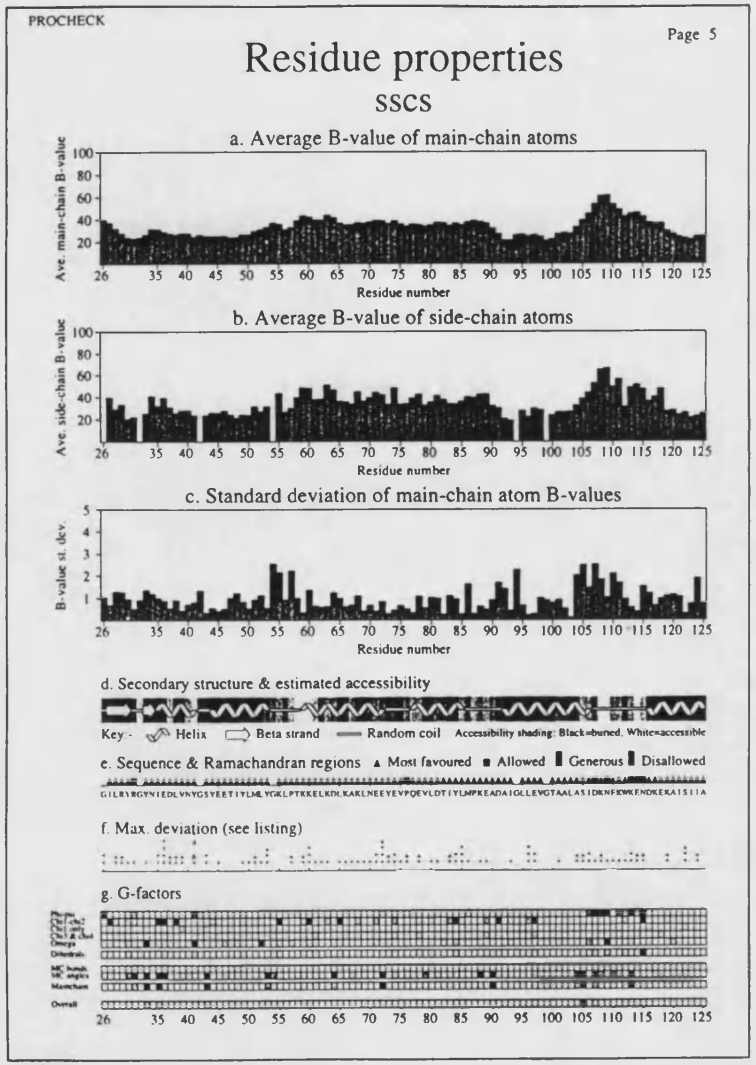




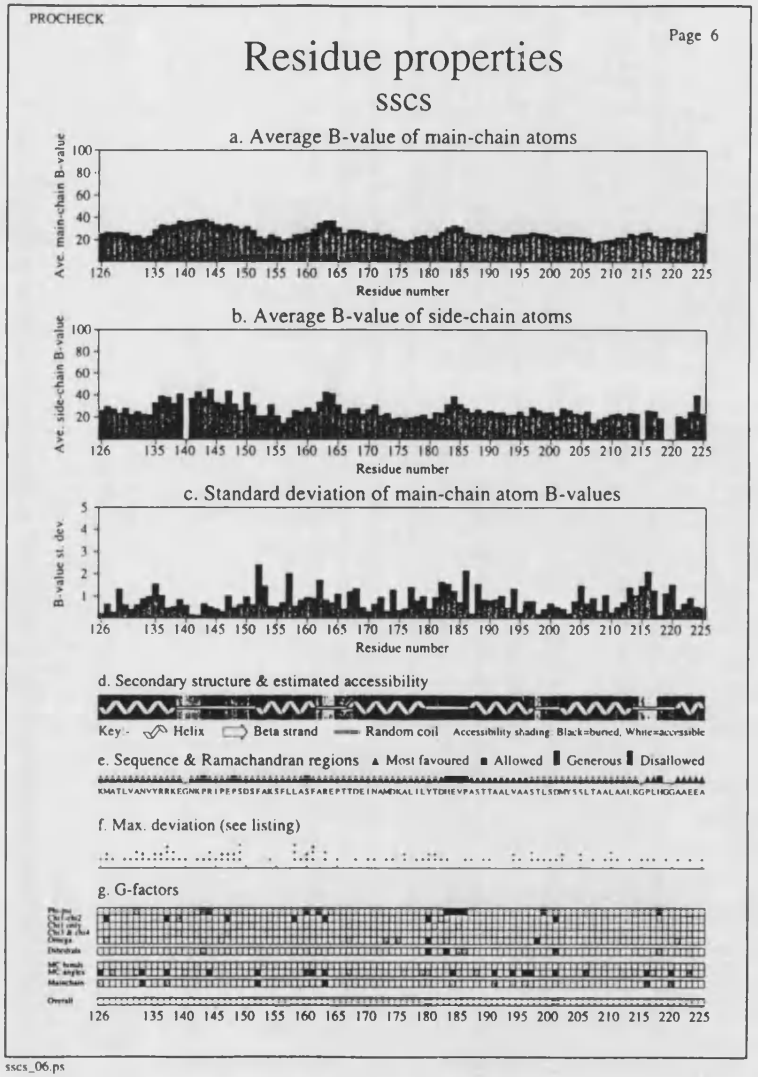
sscs_06.ps



sscs_06.ps



sscs_06.ps

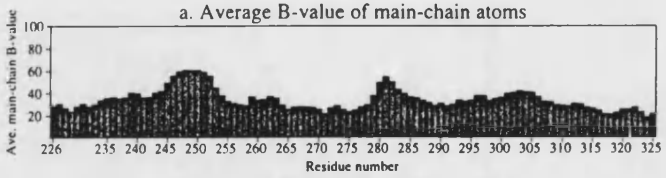


sscs_06.ps

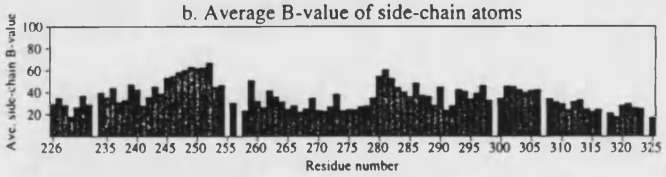
Residue properties

sscs

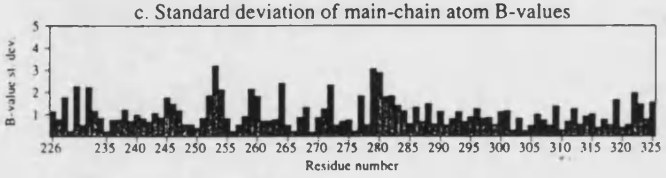
a. Average B-value of main-chain atoms



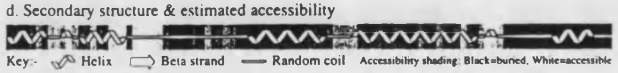
b. Average B-value of side-chain atoms



c. Standard deviation of main-chain atom B-values

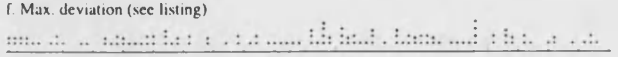


d. Secondary structure & estimated accessibility

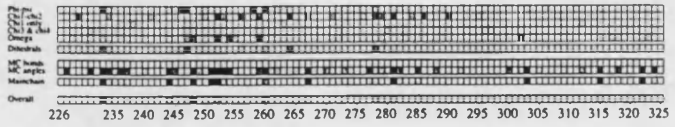


e. Sequence & Ramachandran regions ▲ Most favoured ■ Allowed ▌ Generous ▌ Disallowed

f. Max. deviation (see listing)



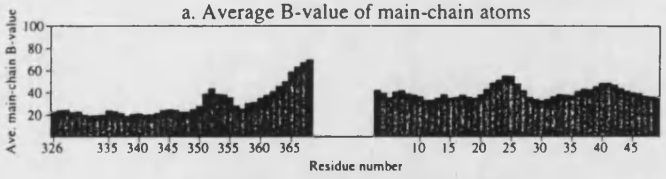
g. G-factors



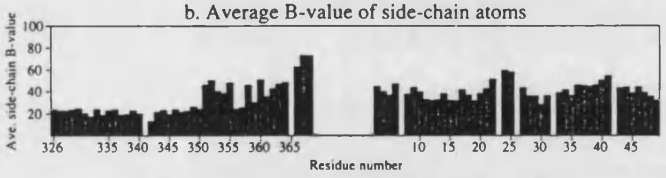
Residue properties

sscs

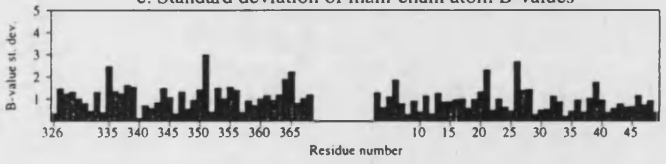
a. Average B-value of main-chain atoms



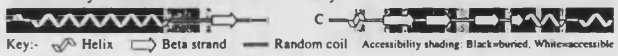
b. Average B-value of side-chain atoms



c. Standard deviation of main-chain atom B-values

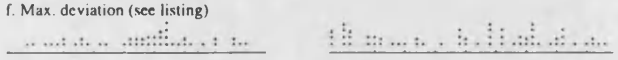


d. Secondary structure & estimated accessibility

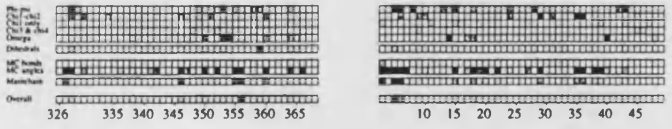


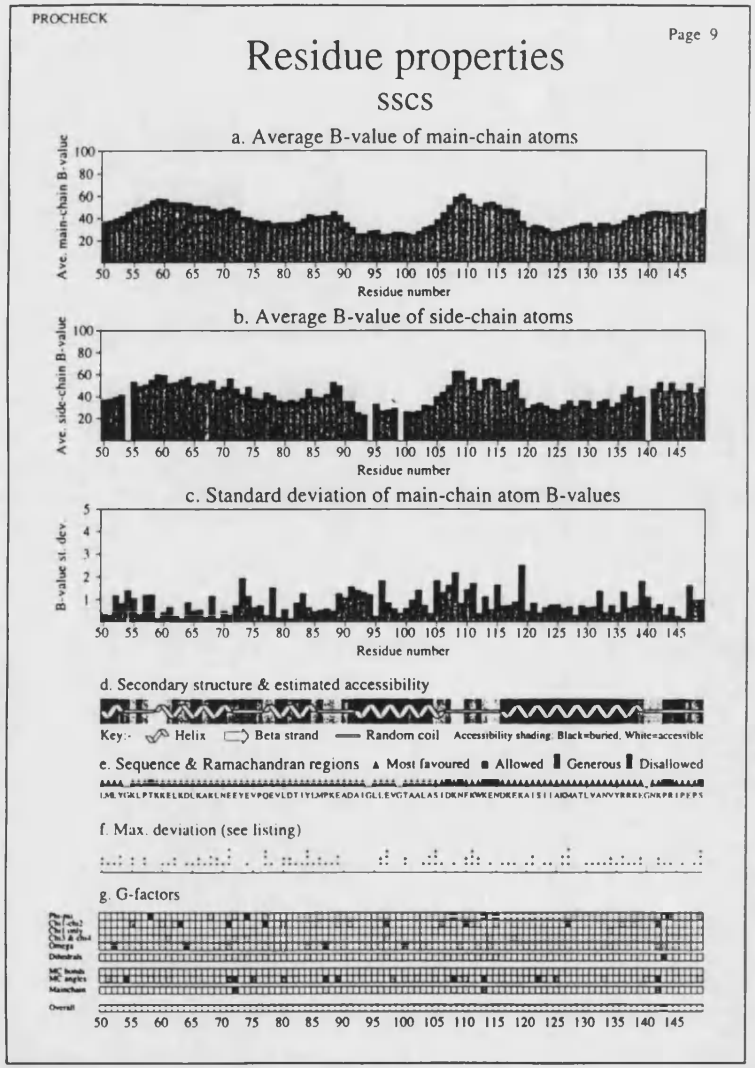
e. Sequence & Ramachandran regions ▲ Most favoured ■ Allowed ▌ Generous ▌ Disallowed

f. Max. deviation (see listing)

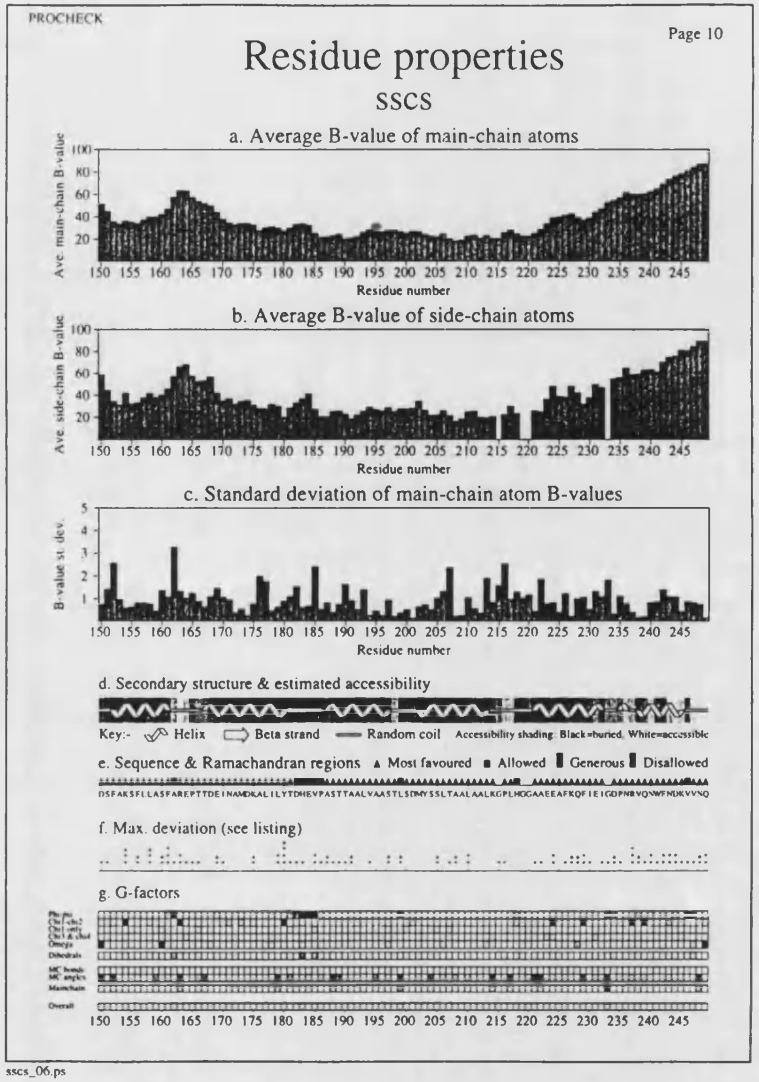


g. G-factors

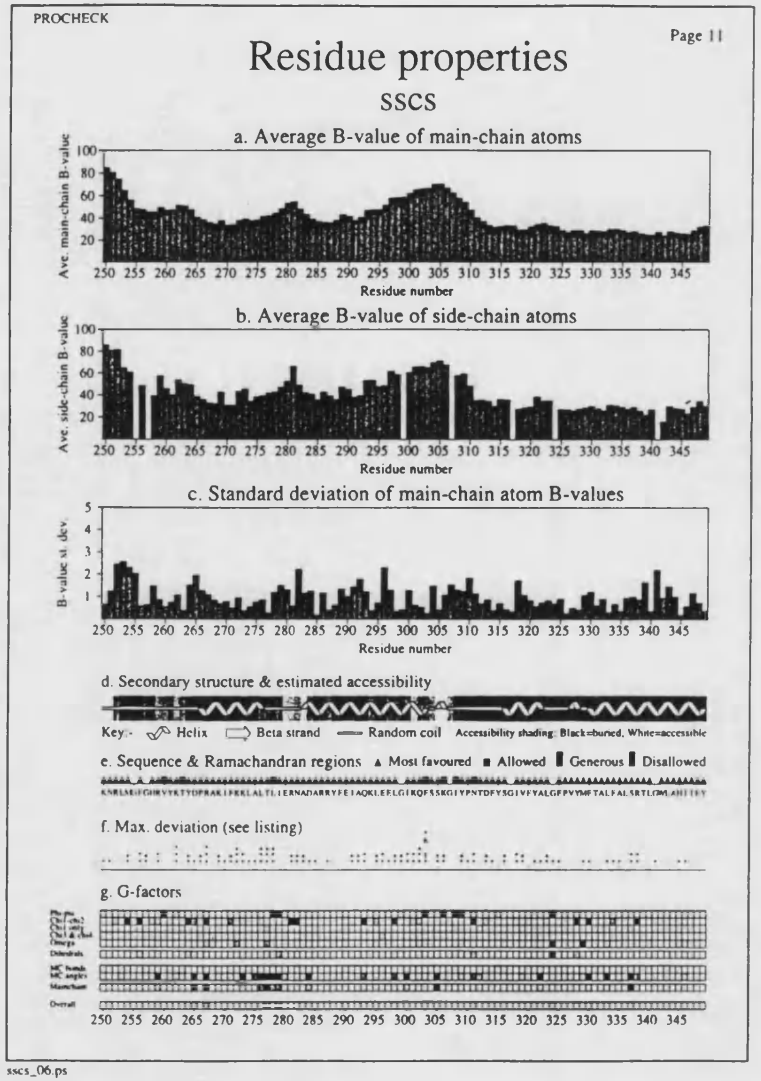




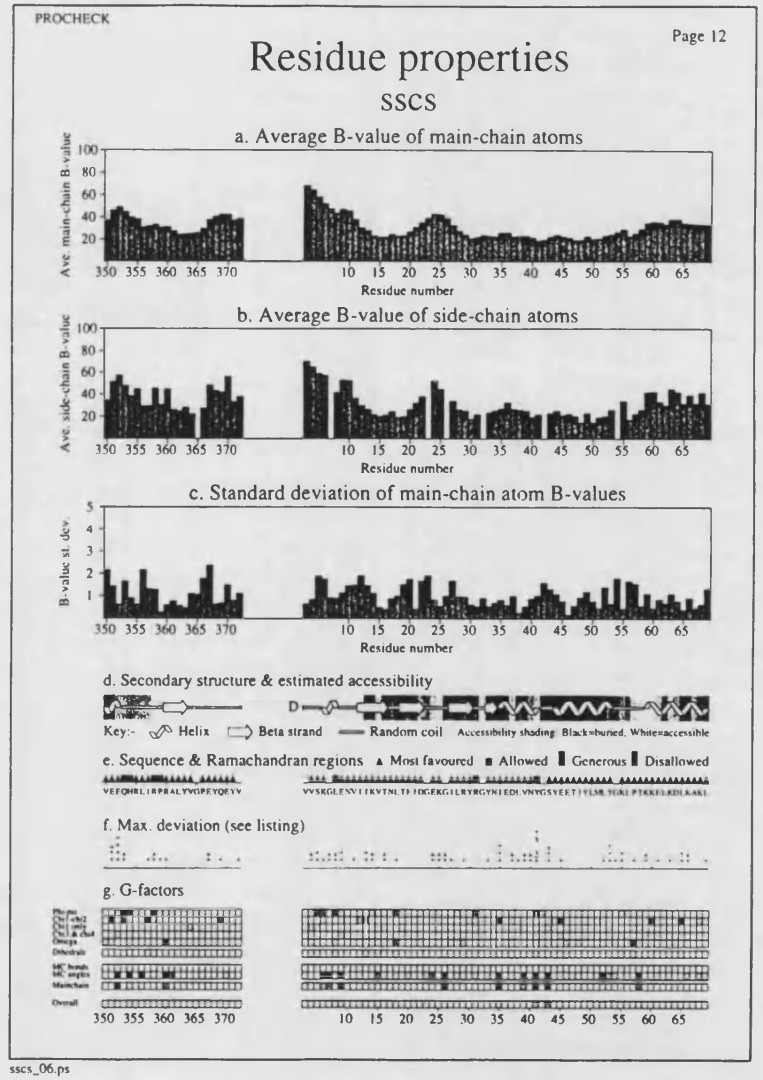
sscs_06.ps



sscs_06.ps



sscs_06.ps

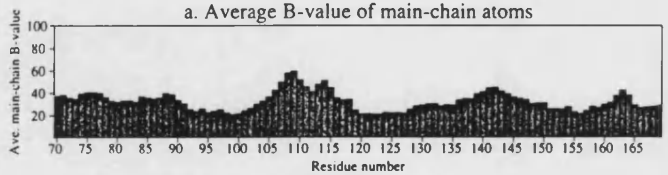


sscs_06.ps

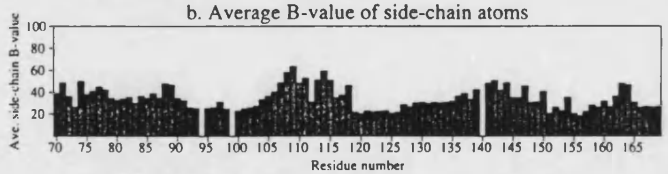
Residue properties

SSCS

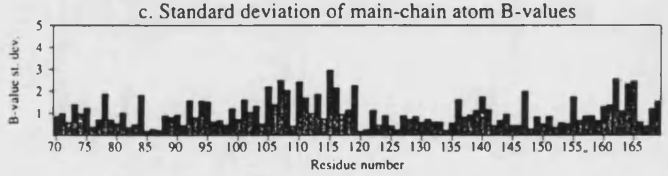
a. Average B-value of main-chain atoms



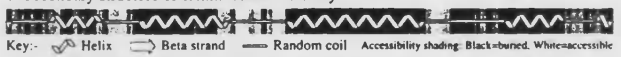
b. Average B-value of side-chain atoms



c. Standard deviation of main-chain atom B-values

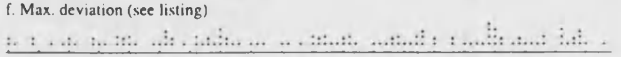


d. Secondary structure & estimated accessibility

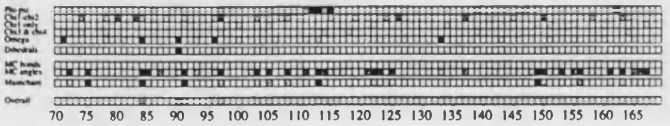


e. Sequence & Ramachandran regions ▲ Most favoured ■ Allowed ▌ Generous ▌ Disallowed

f. Max. deviation (see listing)



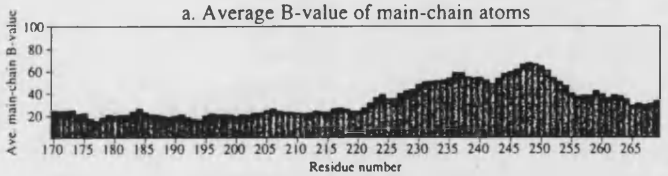
g. G-factors



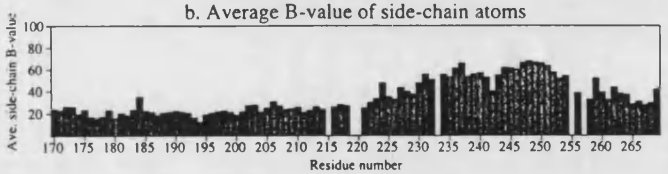
Residue properties

SSCS

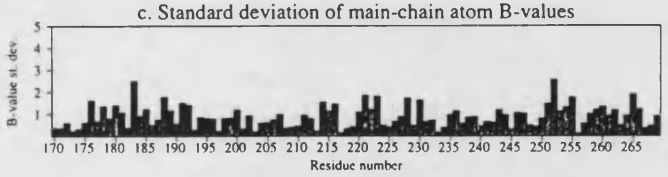
a. Average B-value of main-chain atoms



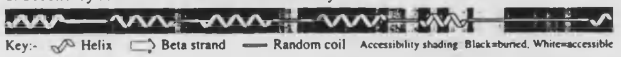
b. Average B-value of side-chain atoms



c. Standard deviation of main-chain atom B-values



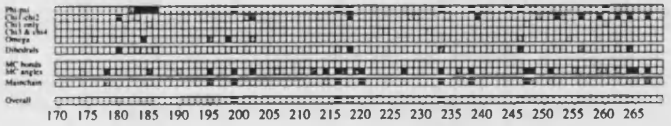
d. Secondary structure & estimated accessibility

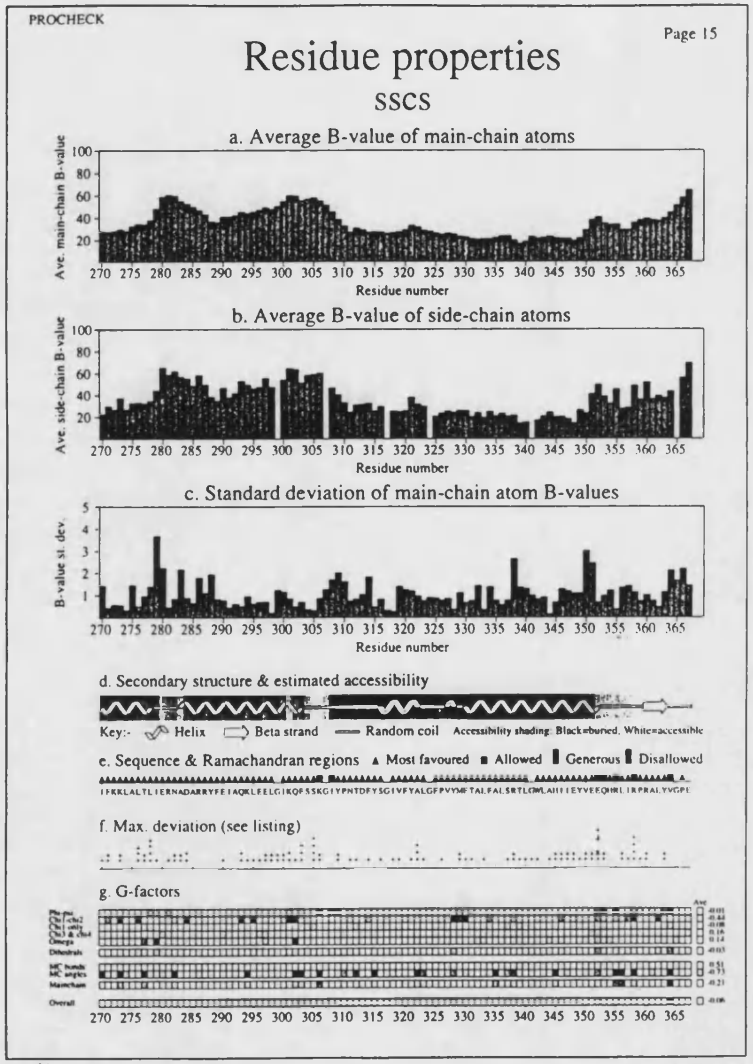


e. Sequence & Ramachandran regions ▲ Most favoured ■ Allowed ▌ Generous ▌ Disallowed

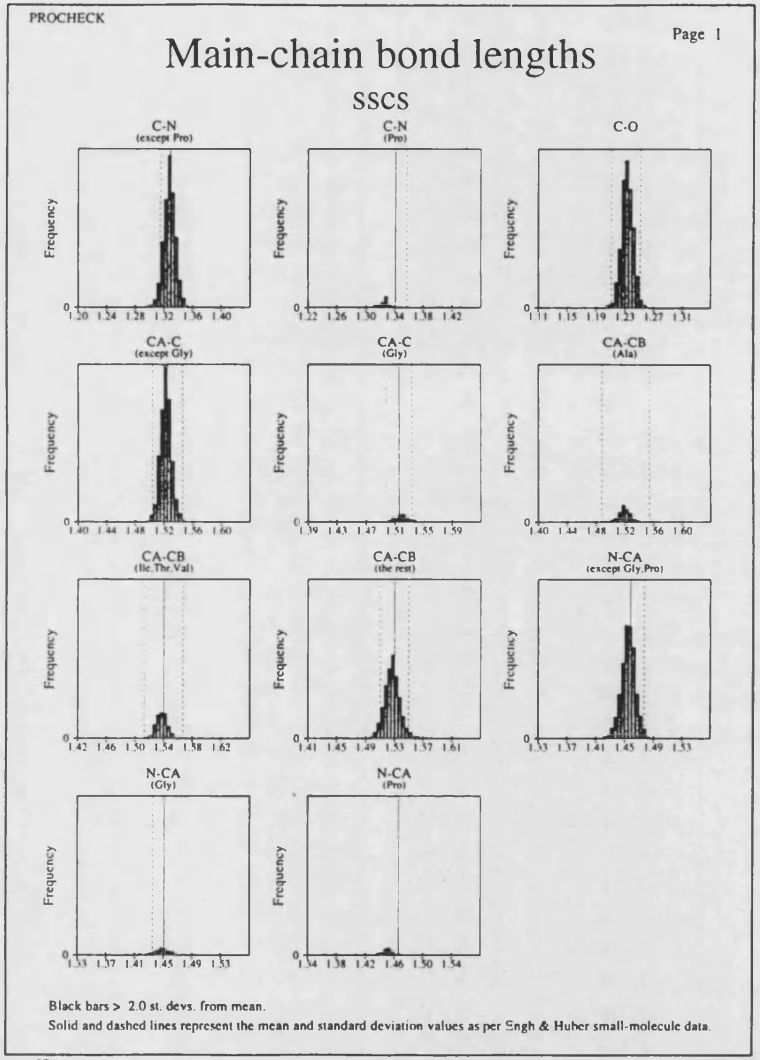
f. Max. deviation (see listing)

g. G-factors





sscs_06.ps



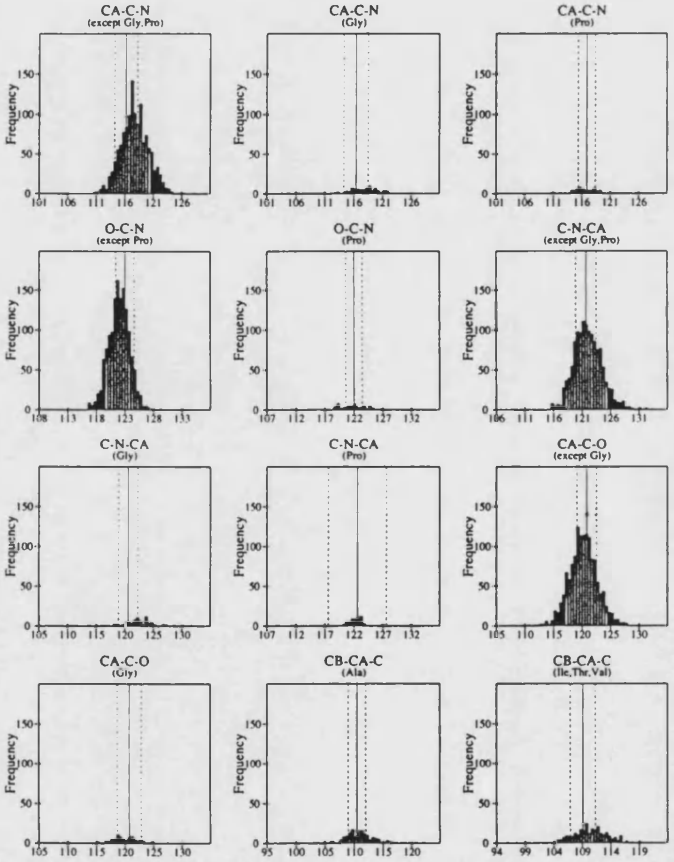
sscs_07.ps

PROCHECK

Page 1

Main-chain bond angles

sscs



Black bars > 2.0 st. devs. from mean.

Solid and dashed lines represent the mean and standard deviation values as per Engh & Huber small-molecule data.

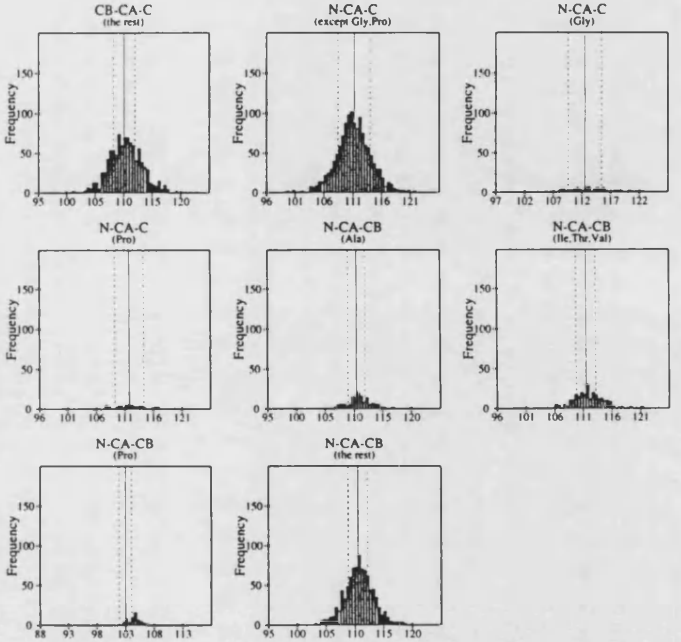
sscs_08.ps

PROCHECK

Page 2

Main-chain bond angles

sscs



Black bars > 2.0 st. devs. from mean.

Solid and dashed lines represent the mean and standard deviation values as per Engh & Huber small-molecule data.

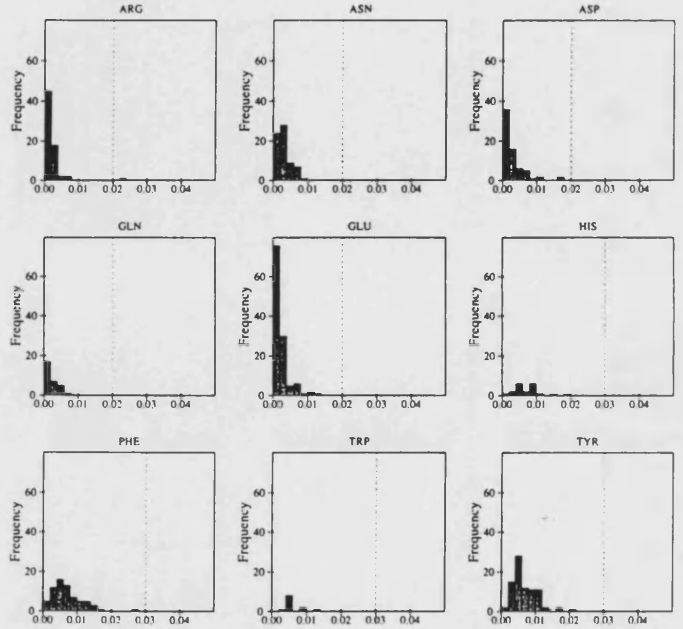
sscs_08.ps

PROCHECK

Page 1

RMS distances from planarity

sscs



Histograms showing RMS distances of planar atoms from best-fit plane.
Black bars indicate large deviations from planarity: RMS dist > 0.03 for rings, and > 0.02 otherwise.

sscs_09.ps

PROCHECK

Page 1

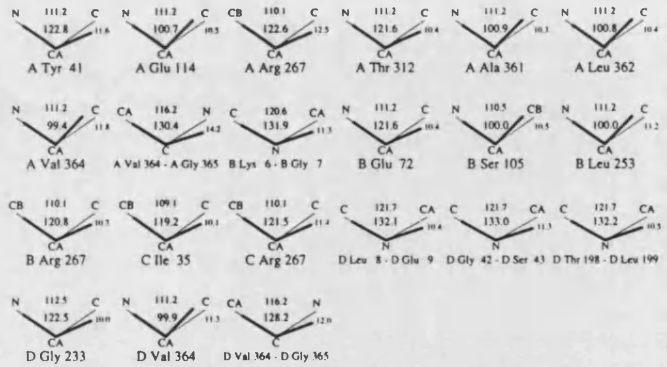
Distorted geometry

sscs

Main-chain bond lengths

Bonds differing by > 0.05A from small-molecule values. Values shown: "ideal", difference, actual

Main-chain bond angles



Bond angles differing by > 10.0 degrees from small-molec values. Values shown: "ideal", actual, diff.

sscs_10.ps

Figure 6.9: Model validation statistics (PROCHECK).

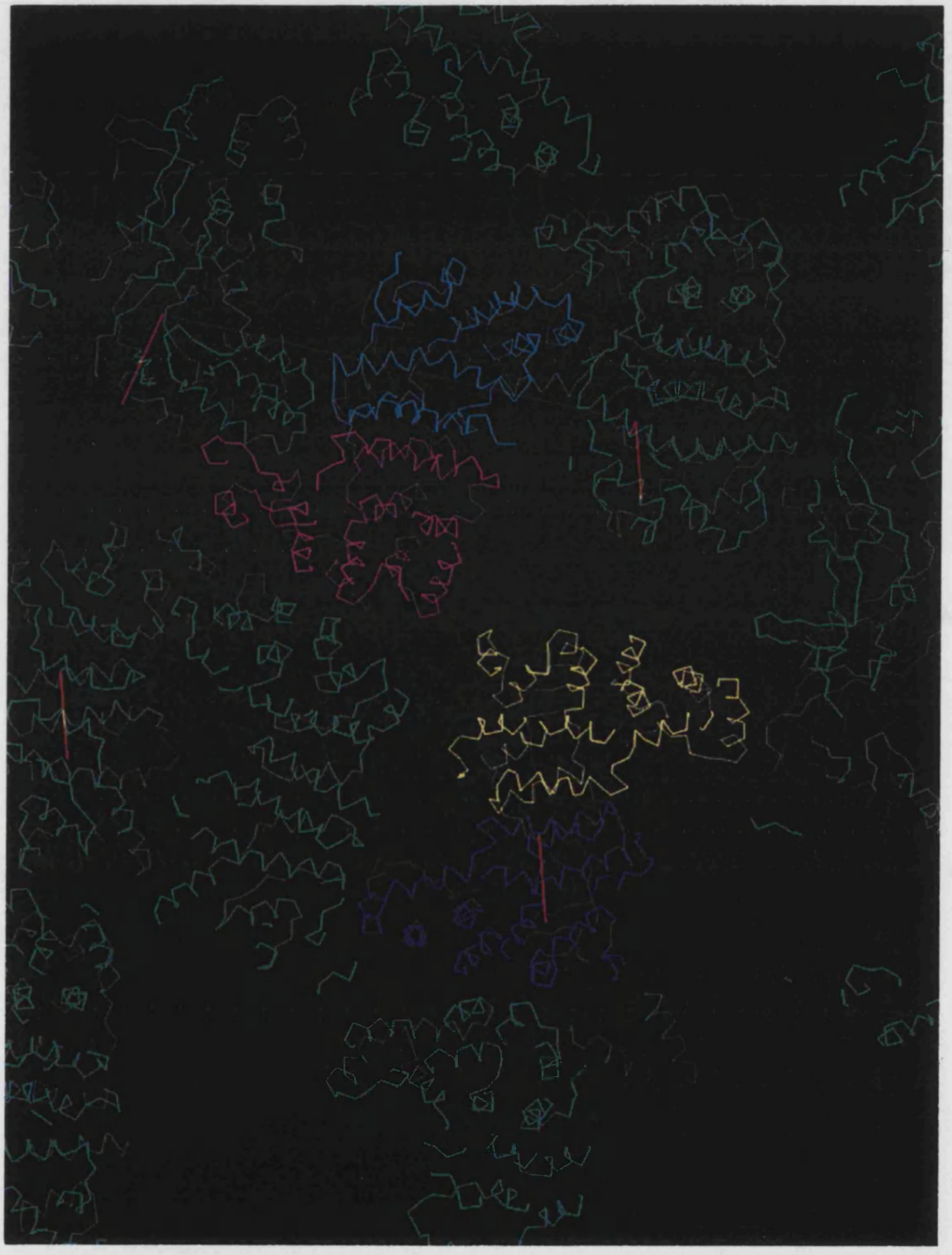


Figure 6.10: Diagram displaying the crystal packing arrangement within the *Ss*CS crystal. The four NCS related monomers are coloured; A is yellow, B, blue, C, cyan and D is magenta).

Figure 6.11: Ribbon diagram showing two orientations of the *Ss*CS dimer with large domains of each monomer coloured light and dark orange, small domains light and dark yellow, and catalytic residues shown in ball-and-stick representation (structure described in the following chapter). The diagram was created using **MOLSCRIPT**.

Figure 6.12: Schematic representation of the SsCS dimer with the α-helices shown as 'rods' and catalytic residues shown in ball-and-stick representation (diagram created with **BOBSCRIPT**).

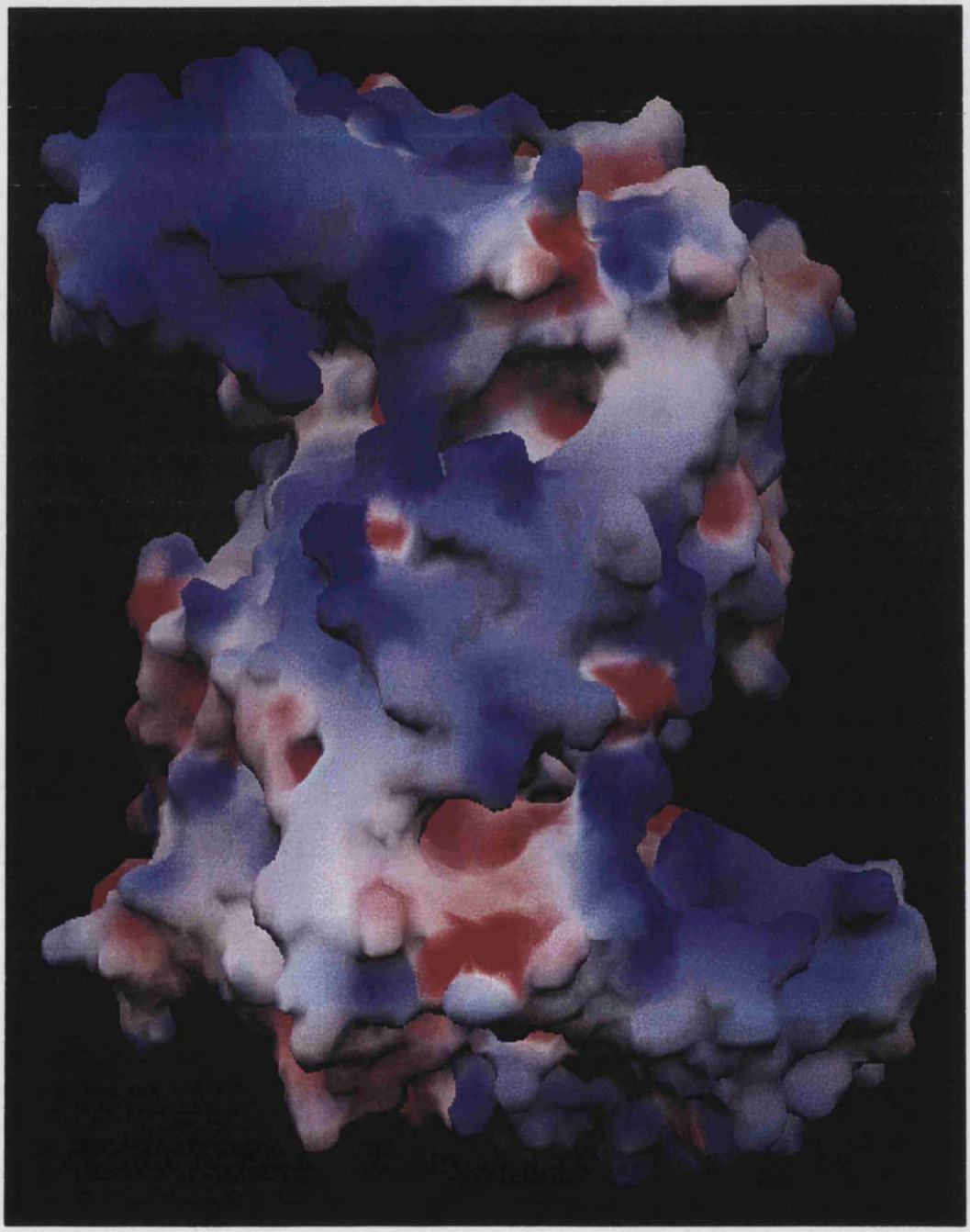


Figure 6.13: Representation of the relative electrostatic potential (red=negative and blue=positive) of the surface of an *Ss*CS dimer demonstrating relative positive potential of the active site cleft. The diagram was generated with **GRASP** [Nicholls *et al*, 1991].

6.6 Discussion

Data collected from the single crystal of *Sulfolobus solfataricus* citrate synthase extended to a resolution of 2.7Å. The knowledge of the 3-D structure of Citrate Synthase from *Thermoplasma acidophilum* allowed straightforward solution of the Molecular Replacement problem in AMoRe, and the model subsequently refined to a final R-factor of 20.8% (R_{free}=28.5%) using the restrained Maximum Likelihood Refinement technique in the program REFMAC. Analysis of the stereochemistry of this final model shows that it is within acceptable limits for the resolution of the data, and it could now be used to extend the study on the structural basis of thermostability of citrate synthases. The structure itself (shown schematically in figures 6.10-6.13) is described in the following chapter with specific reference to the comparison with the known crystal structures.

CHAPTER 7

Structural Analysis of Citrate Synthases; Possible Determinants of Thermostability

7.1 Introduction

Six citrate synthase crystal structures from five host organisms (figure 7.1) have been used for comparative analysis in order to identify some of the structural features which could confer (hyper)thermostability in this enzyme 'family'. Many of the features observed in other proteins have been discussed in the introduction. As can be seen from figure 7.1, the host organisms span the range of temperatures at which life is known to exist. However, in order to qualify the study, it is important to identify the actual inherent stability of each CS *in vitro*, from measured half lives of inactivation [Muir *et al*, 1995, Connaris *et al*, 1998 & Gerike *et al*, 1997]. Care is required when interpreting half-life activities as these are highly dependent on protein concentration, pH and presence of substrates. However, results displayed in figure 7.1 suggest that there may not be quite as large a gap in thermal stability between *Ta*CS, *Ss*CS and *Pf*CS as would be be predicted simply from differences in optimum growth temperature of the organisms.

One of the obstacles to overcome when studying structural differences, is to separate features which actually correlate to differing thermostabilities from those which are merely phylogenetic traits. In the case of the Antarctic bacterium DS23R, there may also be features present which relate more to the problem of overcoming cold denaturation and retaining high enough turnover at very low temperatures. The coupling of observations from crystal structures with SDM studies can help to identify the most important interactions, and results from these studies are mentioned where applicable.

Before embarking on a full analysis of the six crystal structures it is also of fundamental importance to consider limitations of the study. Firstly, the enzymes from each organism have been crystallised either with or without the

presence of substrates (both only in the case of pigCS). As a result of the conformational changes which take place on binding of substrates, only 'true' comparisons can be made between the *Ss*CS, *Ta*CS and pigCS open structures or the *DS*CS, *Pf*CS and pigCS closed structures. These structures have also been refined at differing resolution which may effect the accuracy of measurements of interactions such as ion-pairs and hydrogen bonds (although these may be calculated with differing distance cut-offs in an attempt to account for this). Despite these difficulties an attempt to measure some of the properties of these enzymes has been made. The water structures could not be compared due to lack of high resolution data for the *Ss*CS, *Ta*CS and open pigCS structures.

Source	PDB code	Opt. Growth Temp./°C	t1/2=8mins /°C	Substrates	Data Res.
DS23R (*DS*CS)	—	31*	45	citrate & CoA	2.1Å
Pig (pigCS)	1CTS/ 2CTS	37	58	citrate only/ citrate & CoA	2.7Å/ 2.0Å
T. acidophilum (*Ta*CS)	—	55	87	—	2.5Å
S. solfatricus (*Ss*CS)	—	87	95	—	2.7Å
P. furiosus (*Pf*CS)	1AJ8	100	100	citrate & CoA	1.9Å

Figure 7.1: Citrate synthase structures used for analysis. t1/2=8mins is the temperature at which the enzyme half life of thermal deactivation is equal to eight minutes. *It should also be noted that although DS23R was isolated from a habitat temperature of approximately 0°C, this organism displays a relatively high optimum growth temperature, and thus although it is described here as psychrophilic, should perhaps more correctly be referred to as psychrotolerant or cold-active.

7.2 Overall Comparison of Sequence and Structures

7.2.1 Crystal Structures

All the Eukaryal, Archaeal, gram positive Bacteria are homo-dimeric structures with each identical monomer consisting of a large and small domain. In addition, they are almost entirely α-helical; the pigCS structure contains 20 α-helices (A-T) with *DS*CS, *Ta*CS, *Ss*CS and *Pf*CS containing 16 α-helices, all of which have an equivalent in pigCS (without the extra helices; ABHT). The pig enzyme is therefore the largest enzyme (437 residues) with the sizes of the other enzymes being *DS*CS 379, *Ta*CS 384, *Ss*CS 379 and *Pf*CS 376. Of the 16 equivalent helices, the large domain comprises 11 helices (C-M and S) with 5 in the small domain (helices N-R). The small domain has been classed as residues 217-321 inclusive for *Ss*CS and corresponding residues from the structure based sequence alignment for the other CSs (figure 7.6).

One of the main differences in size of the enzymes is due to the fact that pigCS contains an N-terminal extension of approximately 35 residues containing two helices, of which helix A extends over the surface of the monomer to which it belongs. The pig enzyme also contains a region of antiparallel beta sheet (residues 56-69) which is present in the other CSs (residues 19 to 31 in SsCS). The other CSs also contain an additional region of antiparallel beta sheet brought about by an intersubunit contact of residues 12-16 at the N-terminus with residues near the C-terminus 360-363 (residue numbers correspond to *Ss*CS). Several 3_{10} helices are also present in the SsCS structure; residues 6-8, 235-237, 304-306 and 327-329.

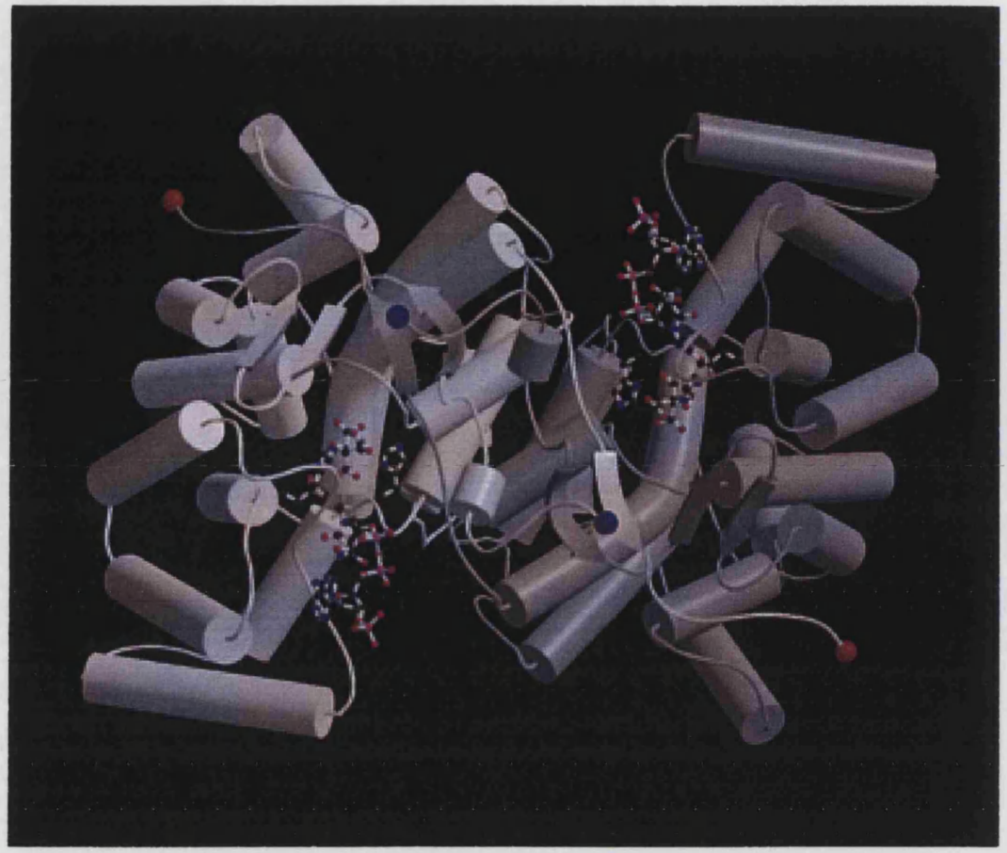
The active site comprises residues from both monomers and therefore CS is only active as a dimer. Active site citrate and CoA binding has been discussed in detail [Russell *et al* 1997 & 1998] for *Pf*CS and *DS*CS and the differences with respect to the pig enzyme. The *Ss*CS structure has no substrate bound but the location of active site residues can be compared. The citrate binding residues comprising three arginine residues; R267 (helix P), R338 (helix S) and R358' (where the prime denotes the residue of the second monomer) and three histidine residues ; H183 (loop K-L), H218 (loop M-N) and H258 (loop O-

P) are equivalent to those found in *Pf*CS. The binding residues for the three phosphate groups of CoA are likely to be K250 (loop O-P) and K306 (loop Q-R), R259 and K262 (both loop O-P), with the third phosphate being coordinated by R355' from the second monomer. The catalytic residues H218, H258 and D313 (loop Q-R) are also present and in a similar position to the *Ta*CS residues. It is likely therefore that *Ss*CS binds substrates in a similar manner to *Pf*CS and carries out the expected acid-base catalysis.

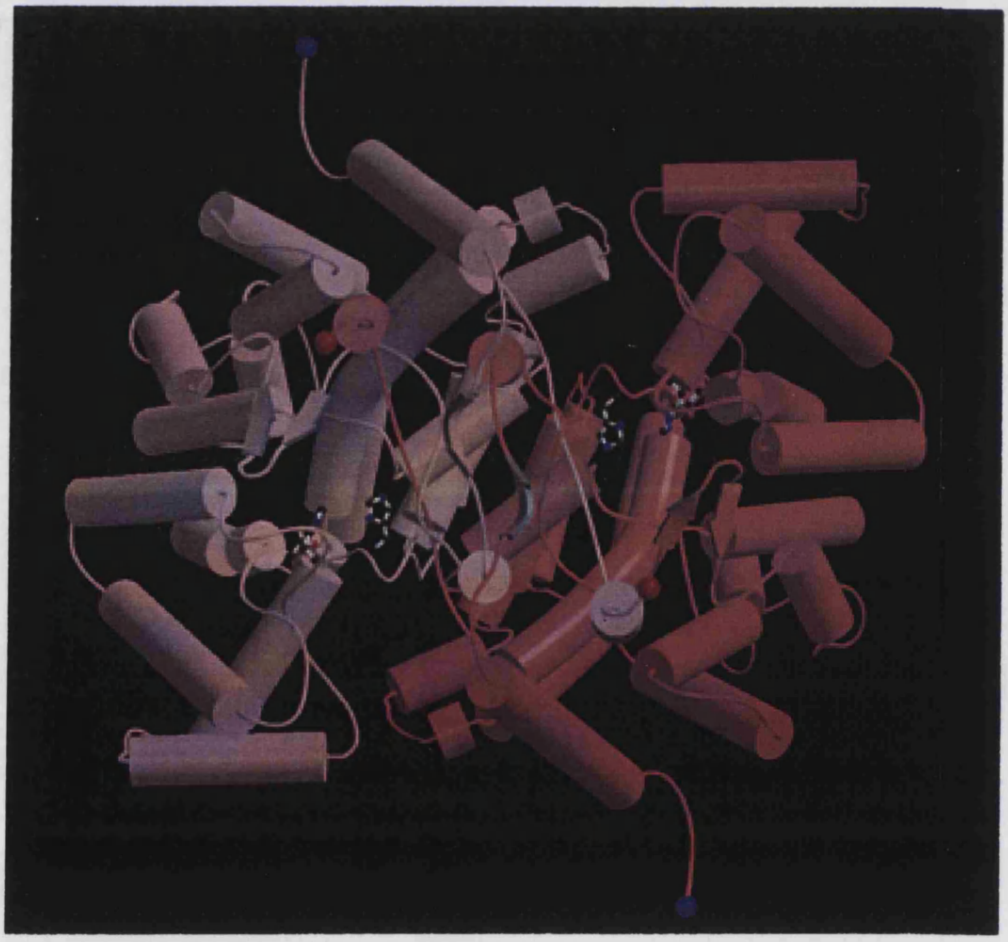
The dimer interface of all the CSs is made up of two parts; the main part being the eight α-helical sandwich of four antiparallel pairs of helices (F,G,L & M), with the additional interaction of N and C-terminal regions. The pigCS is different from the other four citrate synthases in terms of the topology of the C-terminal region. In the other four, the C-terminal arm of one monomer "wraps around the other monomer, clasping the two together in an intimate embrace" [Russell *et al*, 1997] resulting in more extensive interactions, including those with the N-terminus. It is important to note that as the C-terminal arms of the *Ta*CS and *Ss*CS are not complete in the models (below), there may be additional interactions present which have not been observed. This also suggests that the C-terminal arm seems to be ordered only in the presence of substrates.

For *Ta*CS and *Pf*CS, a crystallographic dimer was used in the analysis. For *DS*CS and the open and closed pig CS, both monomers were identical as there is only one monomer in the crystallographic asymmetric unit, with the second generated by a crystallographic two-fold axis. Although the *Ss*CS data is relatively low resolution, there are four monomers on which to base comparison, and after which an average result was generally given. Also, the last 14 residues (the whole of the C-terminal arm) of *Ta*CS and the last 8 residues of the *Ss*CS C-termini were not seen in the electron density map. This will affect many of the results and has had to be taken into account particularly when analysing dimer interface interactions and ion pairs present.

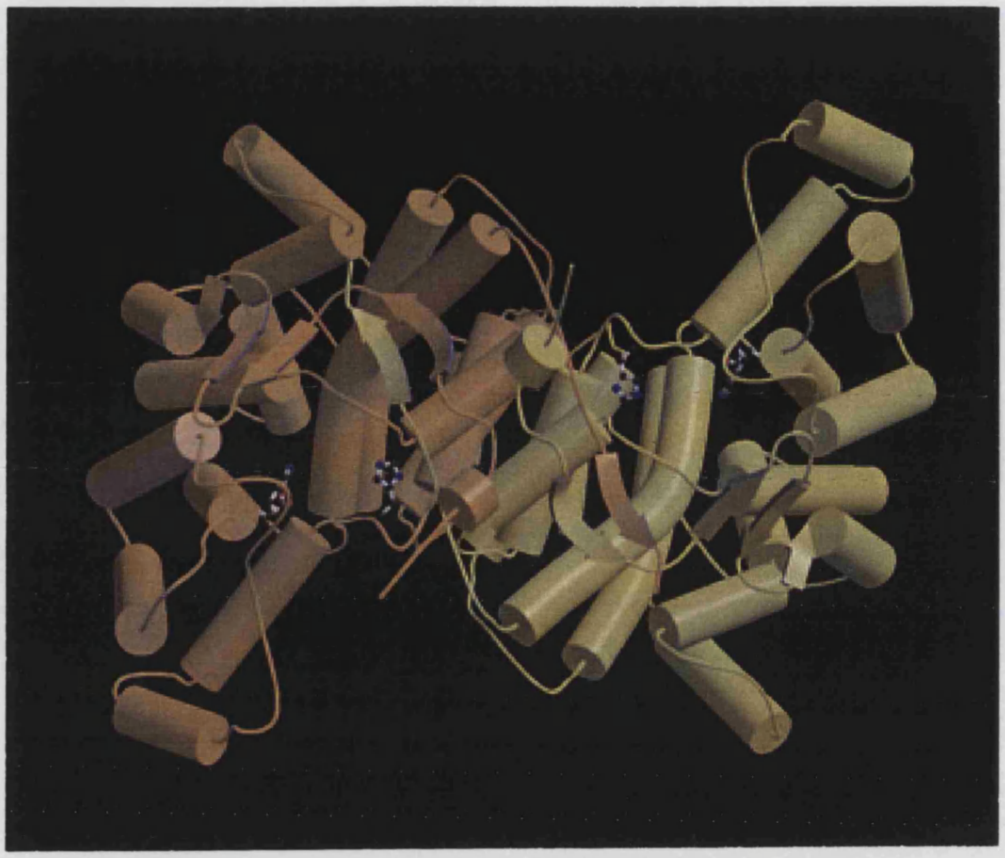
a)



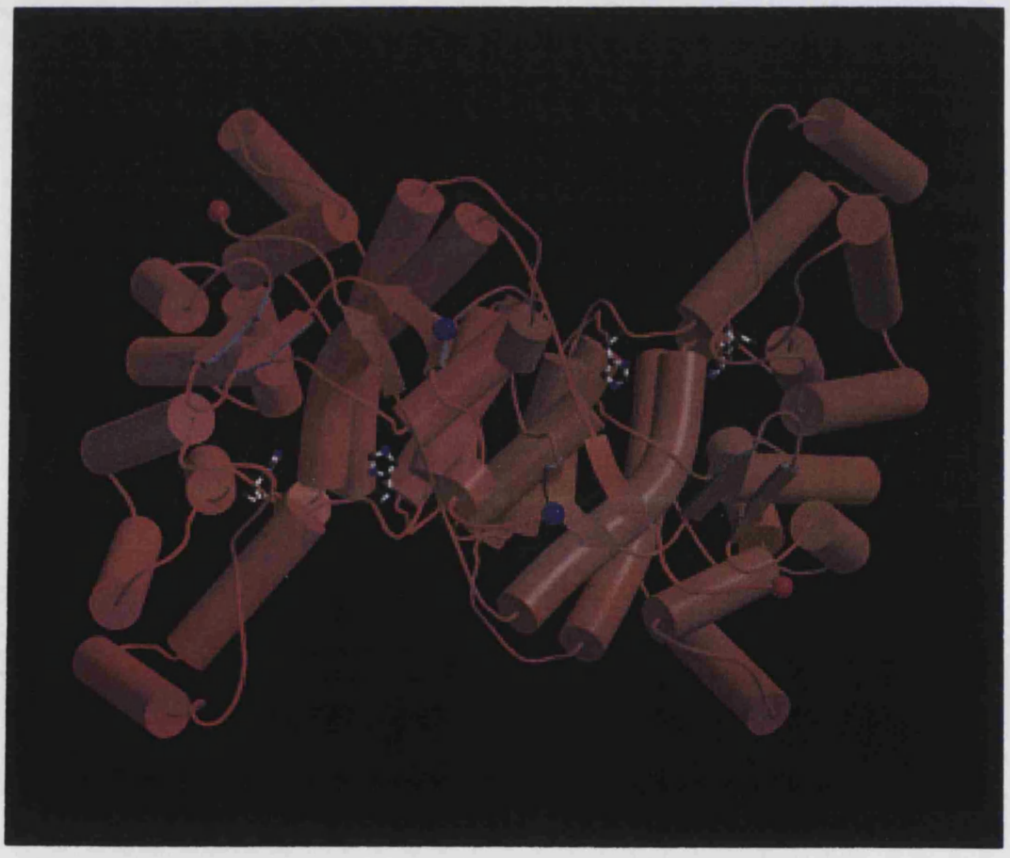
b)



c)



d)



e)

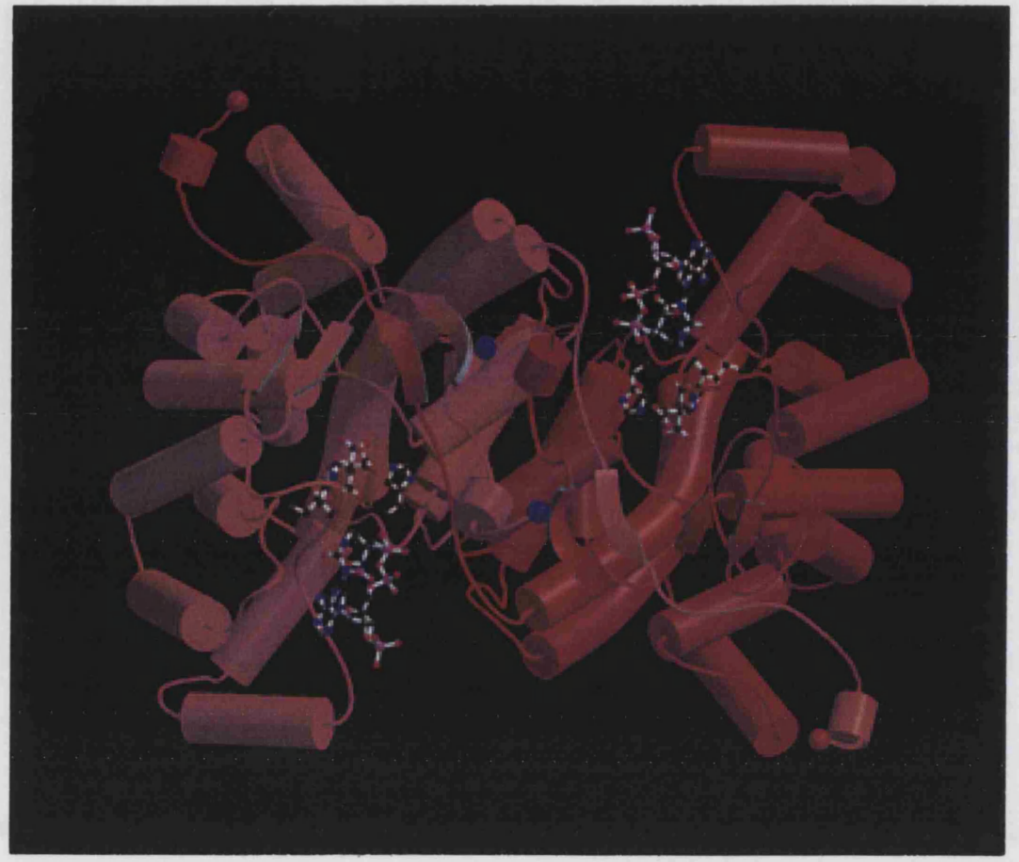


Figure 7.2: Schematic representations of the five citrate synthase structures, with α-helices shown as rods, catalytic residues and substrates shown in ball-and-stick representation (a)*DS*CS; grey/white, b)pigCS; pink, c)TaCS; yellow, d)SsCS; orange, and e)PfCS; red).

Pairwise sequence alignments were carried out using the program **BESTFIT** from the Wisconsin **GCG** sequence analysis package, and superposition carried out using least squares fit in **O** for fitting of alpha-carbon atoms (starting from 3 conserved atoms). These statistics are listed in figures 7.3-7.5. The structure based sequence alignment is shown in figure 7.6.

	DS23R	Pig (open)	Pig (closed)	*Ta*	*Ss*	*Pf*
DS23R	—	2.27 for 560 atoms	2.12 for 630 atoms	1.97 for 604 atoms	1.94 for 610 atoms	1.32 for 719 atoms
Pig (open)	—	—	1.19 for 730 atoms	1.95 for 651 atoms	1.88 for 646 atoms	2.15 for 550 atoms
Pig (closed)	27% (50%)	—	—	2.16 for 533 atoms	2.08 for 519 atoms	2.04 for 631 atoms
Ta	32% (54%)	—	22% (48%)	—	0.87 for 719 atoms	2.03 for 581 atoms
Ss	34% (55%)	—	27% (50%)	59% (76%)	—	1.94 for 597 atoms
Pf	40% (58%)	—	31% (53%)	42% (62%)	46% (67%)	—

Figure 7.3: Overall comparison of primary and 3-D structure of CS dimers with RMS distances (in Angstroms) in the top half of the table, and sequence percentage identities shown in the bottom half of the table (numbers in brackets are similarities).

The Archaeal and Bacterial CSs share a moderate sequence identity (20-30%) with the pigCS and higher identity with each other (with by far the most homologous pair being *Ss*CS and *Ta*CS having a sequence identity of 59%). These sequence statistics are reflected in the RMS deviations between the structures; the Bacterial and Archaeal CSs show a much lower RMS deviation when compared with each other than when compared with the pigCS. The most similar structures are the *Ta*CS and *Ss*CS, with the *Pf*CS and *DS*CS pair also showing very low RMS deviation. As some structures are in the open conformation and some have substrates bound, the large and small domains of each enzyme were compared separately.

	DS23R	Pig (closed)	*Ta*	*Ss*	*Pf*
DS23R	—	1.74 for 242 atoms	1.53 for 252 atoms	1.57 for 259 atoms	1.07 for 262 atoms
Pig (closed)	27.4% (50.4%)	—	1.81 for 233 atoms	1.79 for 232 atoms	1.84 for 245 atoms
Ta	30.1% (52.6)	21.1% (48.4%)	—	0.76 for 256 atoms	1.66 for 247 atoms
Ss	33.5% (55.8%)	28.3% (50.0%)	56.6% (74.4%)	—	1.53 for 245 atoms
Pf	40.8% (58.4%)	30.4% (52.9%)	38.0% (59.5%)	41.1% 62.2%	—

Figure 7.4: RMS distances (Å) and sequence identities for large domains of a single monomer.

	DS23R	Pig (closed)	*Ta*	*Ss*	*Pf*
DS23R	—	1.79 for 96 atoms	1.59 for 90 atoms	1.51 for 90 atoms	1.27 for 92 atoms
Pig (closed)	26.9% (50.0%)	—	1.60 for 82 atoms	1.55 for 82 atoms	1.49 for 90 atoms
Ta	37.8% (59.2%)	24.0% (48.1%)	—	0.76 for 104 atoms	1.02 for 95 atoms
Ss	36.4% (52.5%)	23.1% (51.0%)	64.4% (80.8%)	—	1.03 for 96 atoms
Pf	37.9% (56.8%)	30.6% (52.0%)	54.1% (70.4%)	58.6% (78.8%)	—

Figure 7.5: RMS distances (Å) and sequence identities for small domains of a single monomer.

Separate analysis of large and small domains in general shows the same trend as that for the whole dimer but the small domains are in general more highly conserved. As is suggested later, this may correlate with differences particularly relating to the dimer interface.

7.2.2 Amino-Acid Composition

The information obtainable from amino acid sequence alone gives limited insight into the differences in the thermostabilities of the CSs and therefore obviously needs to be put into a structural context. The differences in the individual amino acids can be seen in the histograms (figure 7.7). Secondary structure allocation was carried out by means of the program **PROMOTIF** [Hutchinson & Thornton, 1996] which uses a local algorithm based on that of Kabsch and Sander [Kabsch & Sander, 1993]. Many of the statistics quoted below are with reference to allocations of secondary structure. Helices were defined as N"N'N_{cap}N1N2N3..... .....C3C2C1C_{cap}C'C" where an internal position was identified as N3-C3 inclusive and helix capping residues N_{cap} and C_{cap} are the residues one outside that defined as helical by the Kabsch and Sander algorithm. A helix initiating position was taken as N_{cap},N1 or N2. Loop regions were defined as C'-N'. In the case where statistics are given for residues in helical or loop environments, only the equivalent helices and loops have been compared.

The so called hot-cold exchanges in the CS structures can be analysed by calculating matrices displaying the 'conversions' for the amino acid which leads to a very complex picture. The analysis carried out here attempts to highlight the more general trends rather than the fate of specific residues.

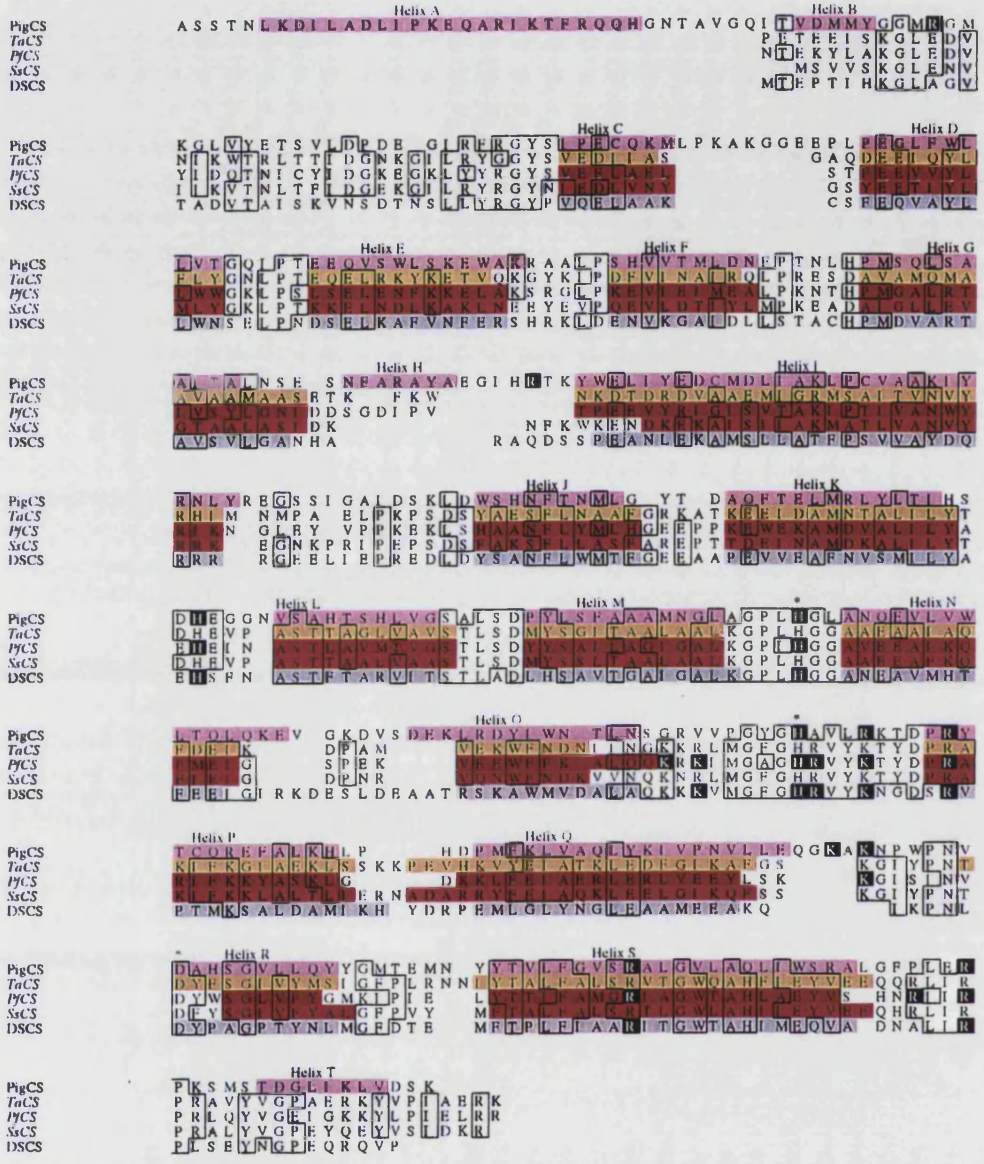
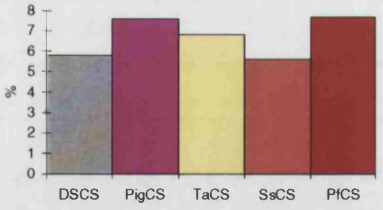
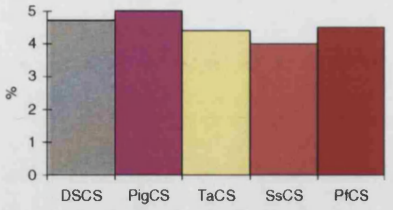


Figure 7.6: Structure based sequence alignment of the five citrate synthases. The helices (A-T) are coloured corresponding to figure 7.2. All identical residues are boxed, the three catalytic residues are marked with an asterisk and the citrate and CoA binding residues are shown in inverse type. The diagram was displayed using **ALSCRIPT**.

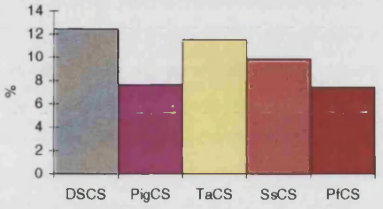
Glycine



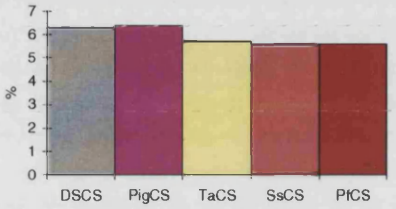
Proline



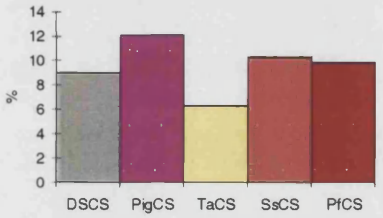
Alanine



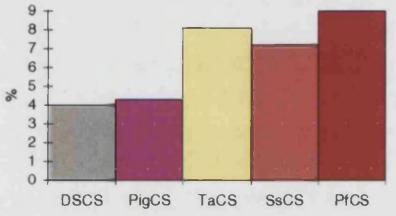
Valine



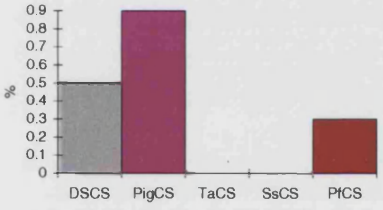
Leucine



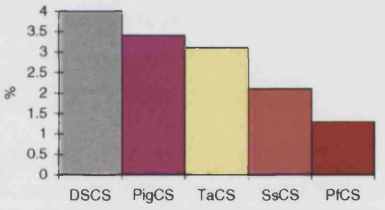
Isoleucine



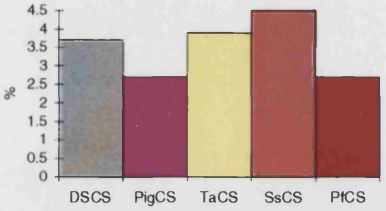
Cysteine



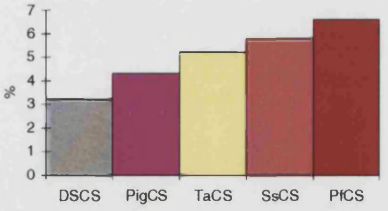
Methionine



Phenylalanine



Tyrosine



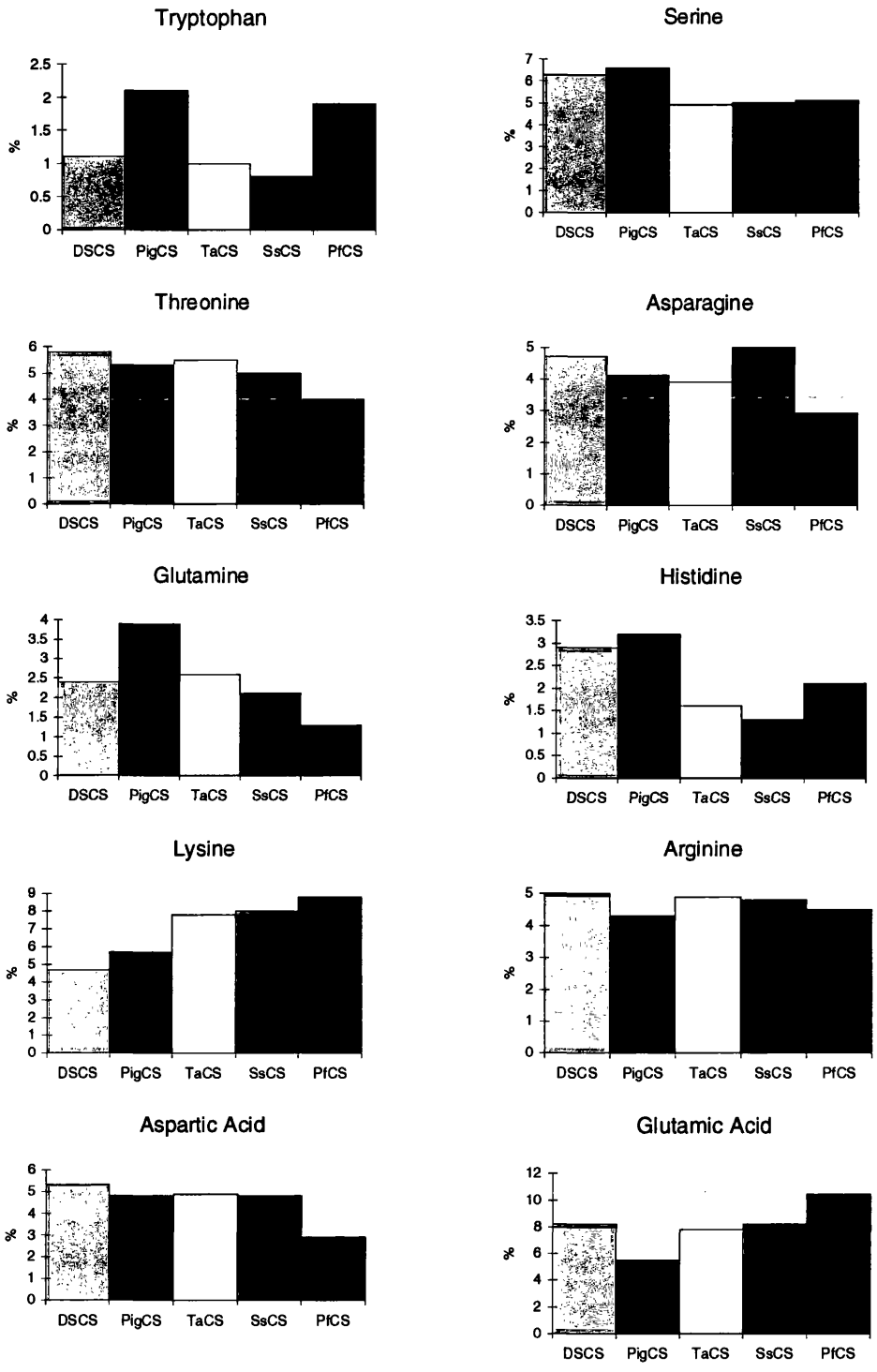


Figure 7.7: Histograms showing the percentage amino-acid composition for the five CSs.

	DS23R	Pig	*Ta*	*Ss*	*Pf*
Charged (DEKRH)	26.1	23.6	27.1	27.1	28.7
Polar/Uncharged (GSTNQC)	25.6	28.4	23.7	22.8	21.3
Hydrophobic (LMIVWPAFY)	48.3	48.0	49.2	50.1	50.0

Figure 7.8: Total percentage contents of charged,polar and hydrophobic side-chains.

In general the thermophilic structures have a higher content of charged residues (although lower content of histidine) than that of the mesophile (although the psychrophile also has a reasonably high charge content). There is also a decrease in the R/(R+K) ratio as you proceed up the temperature scale, which is the converse of that observed in many other proteins; 0.51 for *DS*CS, 0.43 for pigCS, 0.38 for *Ta*CS, 0.37 for *Ss*CS and 0.34 for *Pf*CS. This is mainly due to an increase in lysine content (with all the citrate synthases having a relatively similar arginine content). The higher charge content correlates with the thermophiles having a lower content in polar/uncharged residues. These trends have already been noticed on comparing genome sequences of hyperthermophiles and mesophiles [Deckert *et al*, 1998]. There is not such a notable difference in hydrophobic content (although differences in the amount of hydrophobic residues exposed is discussed later). Amongst the hydrophobic amino acids, isoleucine seems to be favoured in the thermophiles. The result of this is the presence of isoleucine clusters in the thermophiles (section 7.4). There are other noticeable differences such as a slight increase in the content of aromatic residues due to the increase in tyrosine in the thermophiles. Tyrosine should be favoured due to the increase in hydrogen-bonding potential compared with the other aromatic side-chains.

Deamidation of asparagine and glutamine and oxidation primarily of cysteine and methionine is mainly a consideration at hyperthermophilic temperatures and therefore only temperatures at which the *Sulfolubus* and *Pyrococcus* enzymes are likely to operate at. *Pf*CS does have a lower content of both asparagine and glutamine than all the other CSs with an obvious trend towards

discrimination against glutamine in the thermophiles (although *DS*CS also has a relatively low glutamine content). This trend is not present for asparagine. These results correspond with those observed in the genome sequences of *A. aeolicus* and *M. jannaschii* [Deckert *et al*, 1998, Bult *et al*, 1996] in which there was a lower content of glutamine but not of asparagine or methionine when compared with mesophilic genomes. In the citrate synthases there is a decrease in methionine content over all five CSs. *Ta*CS and *Ss*CS also have no cysteine residues compared with two in *DS*CS and four in pigCS. The *Pf*CS sequence contains one cysteine residue. This does not appear to be solvent accessible in the native state of the protein. The thermophiles also have a lower content of histidine (which is also known to undergo oxidation at high temperatures [Daniel *et al*, 1996]) than either pigCS or *DS*CS despite the former organisms having a higher charged residue content. It therefore appears that there is a general decrease in the numbers of thermolabile residues in the thermophiles within this family. The chemical context of the thermolabile residues is perhaps of more significance than simply the total content. Factors affecting deamidation have been discussed [Kossiakoff, 1988] such as hydrogen-bonding of n+1 nitrogen, alpha-carbon backbone conformation and whether the asparagine or glutamine is hydrogen-bonded locally or to a remote part of the polypeptide chain. In the case here no obvious trends could be observed with respect to these chemical factors.

Considering prolines and glycines; there is not a great difference in the overall percentage content of proline in the five organisms, although there are fewer prolines at an internal helix position in the thermophiles and *DS*CS has the lowest number of prolines in turn or loop regions. Although there are comparable numbers of glycines in the five CSs, pigCS certainly has higher number of glycines in loop regions than the other CSs and *Pf*CS also has several more C-capping glycines (C_{cap} C1 and C2) than the other CSs.

	DS23R	Pig	*Ta*	*Ss*	*Pf*
Total no. intra-helical	4	3	0	0	1
Total no. helix initiating	6	7	4	3	5
Total no. in turns/loop regions	8	12	13	12	11

Figure 7.9: Proline residues for comparable regions of secondary structure.

	DS23R	pig	*Ta*	*Ss*	*Pf*
Total no. intrahelical	8	5	7	4	7
Total no. at C-cap of helices	3	4	4	4	7
Total no. in turns/loop regions	7	15	8	6	7

Figure 7.10: Glycine residues for comparable regions of secondary structure.

7.3 Overall stability of alpha helices

Some features (in addition to occurrence of proline and glycine residues mentioned above) which may be associated with the stability of the helices have been included in the table in figure 7.11. Charge dipole statistics refer to stabilised helical dipoles (for example, negatively charged residue at an N_{cap}, N1 or N2 position of a helix) and numbers in parentheses refer to destabilising interactions. On the evidence of helix dipole statistics alone, the pigCS helices may well be more stable than those in the thermophiles. There does not appear to be an obvious trend in the numbers of i(i+3) or i(i+4) ion pairs, which has previously been linked to helix stability, although it is of interest that DS23R has the greatest number. Fewer beta-branched amino-acids (V,I and T) have been found at internal helix positions of thermophilic helices [Facchiano *et al*, 1998] (thought to be due to the extra steric hinderence) but the opposite trend is found in this case.

	DS23R	Pig	*Ta*	*Ss*	*Pf*
Internal beta-branched (VIT)	30	27	32	31	31
Charge dipole(Nter)	7(2)	8(3)	9(0)	7(2)	5(2)
Charge dipole(Cter)	5(3)	6(2)	2(4)	1(5)	6(5)
i(i+3) or i(i+4) salt bridges	8	3	5	3	5

Figure 7.11: Several features associated with stability of the sixteen equivalent α-helices.

7.4 Compactness

The overall accessible surface area (ASA) and volume, and the presence of internal cavities have been calculated, which should be linked to the different packing of the folded proteins. The reduction in the size of loop regions may also be important to this end but has been discussed separately (section 7.8). In addition the different types of area exposed are listed which may be important in minimising surface interactions with solvent as well as maximising buried hydrophobics.

7.4.1 Amino-Terminus

As previously mentioned, one of the main differences in the pig enzyme (compared with Archaeal and Bacterial CSs) is the N-terminus extension. This extension forms a large helical region (helices A and B) which folds over the surface of the enzyme. It is difficult to establish whether this feature is only a phylogenetic distinction, and certainly citrate synthases from some mesophilic organisms seem to be of the shorter type (such as *E. coli* CSII [Patton *et al*, 1993]). For the purpose of calculating some statistics (below) this part of the pig enzyme has been removed.

7.4.2 Accessible Surface Area and Volume

The accessible surface area calculations were calculated using the program **GRASP** and volume calculations using the program **VOIDOO** [Kleywegt & Jones, 1994] with a probe radius 1.4Å and grid spacing of 0.75Å. All calculations for closed structures were done in the absence of substrate.

*DS*CS, *Ta*CS and *Pf*CS all have a very similar surface area, with SsCS slightly higher but all have considerably smaller surface area and volume than the pigCS, even when deleting the first 35 residues from the pig enzyme. A similar pattern to the total ASA is found when comparing the overall volume; pigCS having considerably larger volume than the other CSs (again even when calculating with the N-ter deleted pigCS). However it is also notable that the smallest voulme is exhibited by the psychrophile being only 8.36 x10^4Å^3. All the CSs have comparable percentages of atoms buried (*Pf*CS the highest with 54.5%). Examination of the hydrophobic area exposed shows a more obvious trend. Despite all the Archaeal and Bacterial CSs having a similar overall ASA, there is a considerable difference in hydrophobic exposure when comparing *DS*CS with the other CSs. *DS*CS closed having overall 7854Å^2 exposed hydrophobic area (representing 29%of the total ASA) compared with *Pf*CS closed 4942Å^2 (18%of total ASA). The increase in hydrophobic area exposed in *DS*CS is accounted for by a reduction in the number of exposed lysine residues when compared with the thermophilic CSs. Figure 7.13 demonstrates that *DS*CS on average exposes 23Å^2 per hydrophobic residue compared with 16Å^2 per residue in *Pf*CS. The total amont of hydrophobic area exposed is also greater for the PigCS, *Ta*CS and *Ss*CS than in *Pf*CS, however, when considered as a percentage of the total ASA, perhaps the difference between the *DS*CS and the others is the only significant one. The only open/closed comparison is that of pig and it can be seen that in this case the closed structure has a slightly reduced exposed hydrophobic surface compared to the open structure, showing the fraction of hydrophobic area exposed when substrates are not bound.

	DS2-3R	Pig (open)	Pig (closed)	Pig (closed, Nter deleted)	*T a*	*S s*	*P f*
ASA/x10^4Å^2	2.72	3.34	3.20	2.99	2.72	2.82	2.72
No. of atoms calculated for	5784	6888	6884	6344	5722	5879	5961
No. of atoms buried	3044	3469	3601	3307	2955	3014	3248
Atoms buried (%)	52.6	50.4	52.3	52.1	51.6	51.3	54.5
Volume x10^4Å^3	8.36	9.96	9.98	9.18	8.71	8.51	8.65
Total Exposed Hydrophobic /Å^2	7854	6654	6246	-	6001	5513	4942
% Hydrophobic of total ASA	29	20	20	-	22	20	18
Total Exposed Polar/Å^2	6401	10401	10020	-	5684	7292	5632
Total Exposed K/Å^2	3671	5679	5825	-	5989	5921	6957
Total Exposed R/Å^2	1767	3317	2599	-	2822	2958	1944
Total Exposed D/Å^2	2138	2413	2457	-	2360	1572	988
Total Exposed E/Å^2	4655	3441	3475	-	3371	4743	5974

Figure 7.12: ASA and volume statistics (calculated on a residue by residue basis).

	DS23R	Pig (closed)	Pig (open)	*Ta*	*Ss*	*Pf*
Average Area K/Å^2	102	117	114	111	96	109
Average Area R /Å^2	47	68	87	83	87	59
Average Area D /Å^2	59	58	57	62	46	45
Average Area E /Å^2	86	72	72	68	79	79
Average Area Polar /Å^2	32	38	39	31	41	32
Average Area Hydrophobic /Å^2	23	17	18	19	17	16

Figure 7.13: Average area exposed per residue type calculated for CS dimers using GRASP.

7.4.3 Internal Cavities

Calculations of internal voids were carried out using the program VOIDOO to calculate the probe-occupied cavities with the parameters 1.4Å probe and 0.75Å grid using the multirotational approach described by Kleywegt and Jones [Kleywegt & Jones, 1994] and performed on ten randomly rotated dimers.

The method of calculating cavities may be extremely sensitive to how well the structure has been refined and also the algorithm used (the results differ slightly from those published [Russell *et al*, 1997]), but it does seem evident that there are fewer cavities in the thermophilic *Ta*CS when compared with the open form of the pig enzyme and no cavities were detected in either of the hyperthermophilic proteins. As discussed earlier [Russell *et al*, 1997] the cavity calculated for *DS*CS must be treated with caution as it is bounded by residues from a flexible loop region which was poorly defined in the electron density map.

	DS23R	Pig (open)	Pig (closed)	*Ta*	*Ss*	*Pf*
No. of Internal Cavities	(1)	6	3	3	0	0
Total Cavity Volume/Å^3	(104)	476	218	184	0	0
% Total Volume	(0.1)	0.5	0.2	0.2	0	0

Figure 7.14: Table listing internal cavities calculated with **VOIDOO**.

7.4.4 Isoleucine clusters

As mentioned earlier, the three thermophiles have a higher isoleucine content than either the pigCS or *DS*CS. This results in isoleucine clusters being present in the core of these proteins. In particular, it has already been identified [Russell *et al*, 1997] that one hydrophobic cluster of six isoleucine residues (15,190 and 355 from both monomers) surrounded by a ring of ion pairs is present between intersubunit helices F,G,L and M and helices I and S in each monomer in *Pf*CS. This replaces a four-residue intersubunit network (D420, H246) in pigCS (discussed later). Isoleucine 15 is conserved in both *Ta*CS and SsCS and isoleucine 355 in *Ta*CS, *Ss*CS and *DS*CS. Figure 7.15 shows the distribution of isoleucine residues in the five CSs, with the thermophiles displaying a greater degree of clustering in the protein core (the ionic interactions in *Pf*CS also displayed in figure 7.15b are discussed in section 7.7).

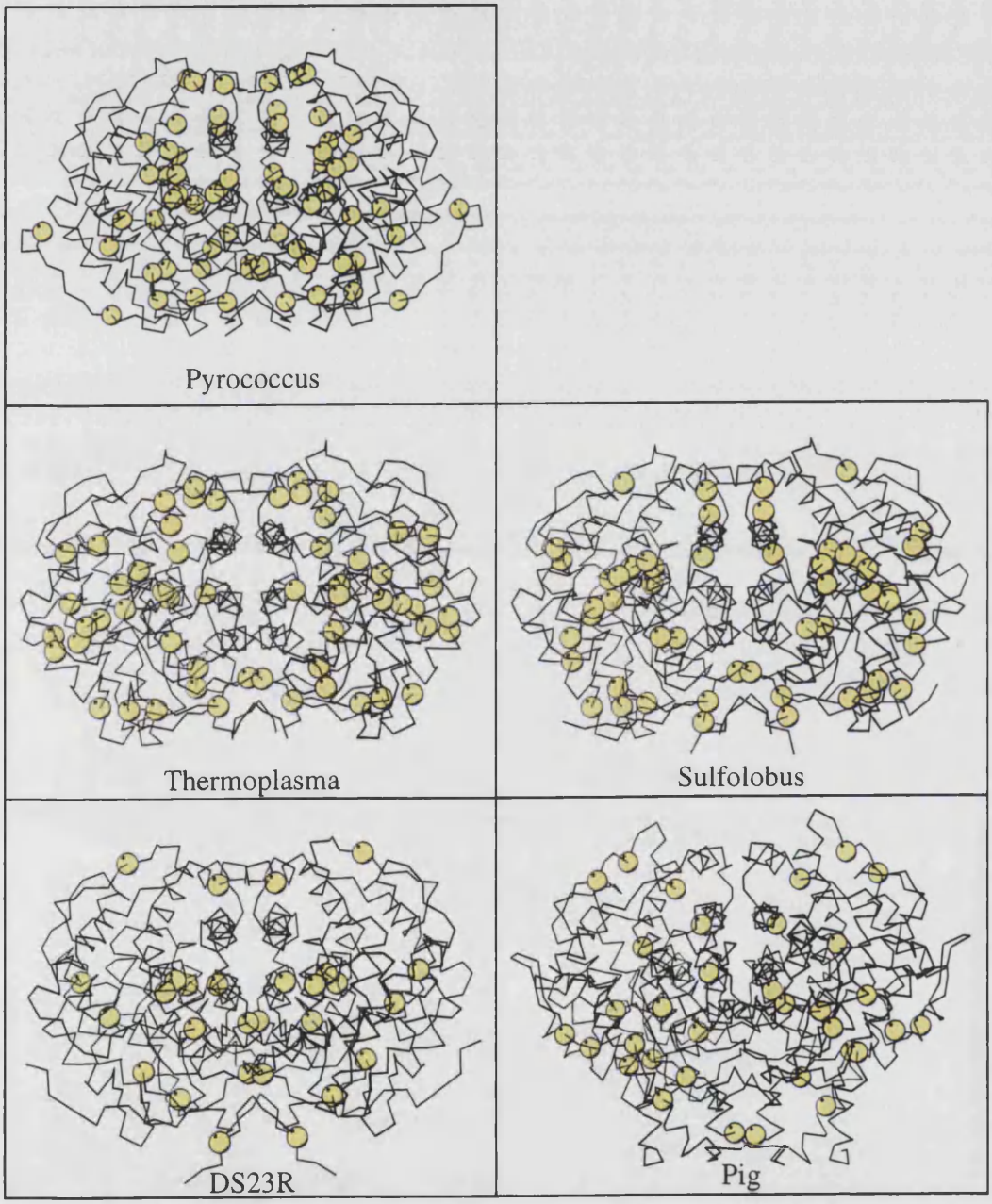


Figure 7.15a: Isoleucine residues in the five citrate synthases (shown as yellow spheres).

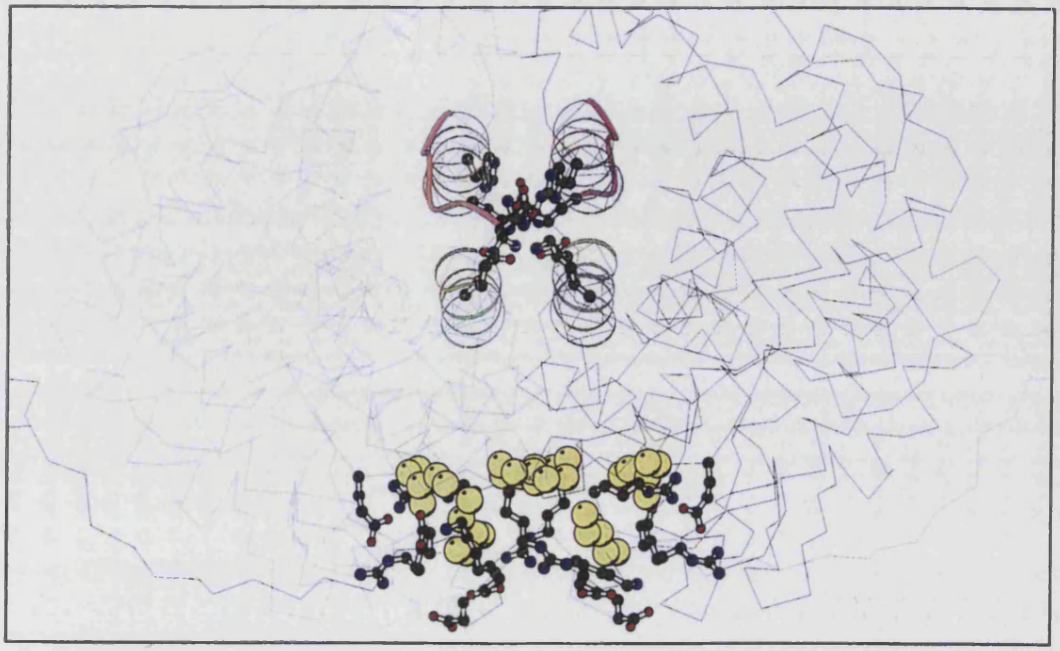


Figure 7.15b: The isoleucine cluster and ionic interactions at the dimer interface in *Pf*CS, with isoleucine residues again shown in yellow.

7.5 Hydrogen Bonding

Hydrogen bonds were calculated for the dimers with the program **HBPLUS** [McDonald & Thornton, 1994]. The default cut-off of 3.5Å was used.

	DS23R	Pig (closed)	*Ta*	*Ss*	*Pf*
Total no. of H-bonds	748	865	745	713	737
No. of H-bonds per residue	0.99	0.99	0.97	0.95	0.98

Figure 7.16: Total number of hydrogen-bonds and number per residue.

It is difficult to tell what role hydrogen-bonding plays in the stability of the citrate synthases. As with ionic interactions, it is likely that the specific location of hydrogen-bonds throughout the structure is more of interest when considering features that are important for thermostability. As can be seen later, there may be differences at specific areas such as the dimer interface.

7.6 Flexibility (B-factors)

Due to the fact that the B-factors depend greatly on the refinement procedure, nature of the crystal (contacts and solvent content) temperature of data collection or resolution of data, they can not be compared empirically between the different structures. Several trends can however be observed. In all cases, the average B-factor of the small domain is higher than that for the large domain. The generally high regions of flexibility have been highlighted for *Ss*CS (previous chapter). This reflects the higher flexibility expected due to the fact that the small domain rotates relative to the large domain on binding substrate. It has already been observed that the relative value is highest for *DS*CS [Russell *et al*, 1998], as seen in figure 7.17 and in the B-factor plots by residue (figure 7.18). It has been suggested that the increased relative B-factor of the small domain of *DS*CS perhaps reflects the greater need for catalytic turnover at low temperatures, but in addition this increased flexibility could also be indicative of a reduced stability of this region in *DS*CS.

	DS23R	Pig (closed)	Pig (open)	*Ta*	*Ss*	*Pf*
Av. B-factor (large domain)	12.3	14.3	22.6	17.4	34.3	21.7
Av. B-factor (small domain)	19.2	16.2	25.1	21.3	45.8	22.9

Figure 7.17: Average B-factors for small and large domains.

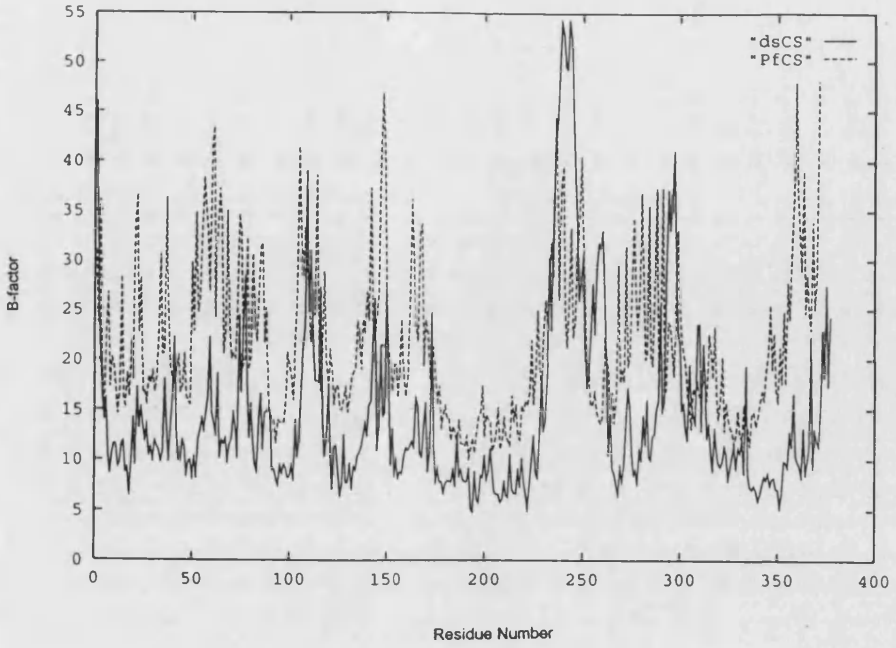


Figure 7.18: B-factor plots per residue for *DS*CS and *Pf*CS (dotted line).

7.7 Ionic Interactions

7.7.1 Ion Pairs

For most comparisons, ion pairs were classed as residues of opposite charge situated 4.0 Å or less apart, [Thornton & Barlow, 1983] but to take into account the flexibility of these residues and errors due to refinement at medium/low resolution (pigCS (open), *Ta*CS, *Ss*CS) then ionic interactions using a 5.0Å and 6.0Å cut-off have also been calculated using a program written by Garry Taylor.

	DS23R	Pig (closed)	*Ta*	*Ss*	*Pf*
Total Ion Pairs	52	36	43	45	43
Intra A	21	12	18	21	14
Intra B	21	12	21	18	12
Inter AB	10	12	4	6	17

Figure 7.19a: Ion pairs, 4.0 Å cut-off.

	DS23R	Pig (closed)	*Ta*	*Ss*	*Pf*
Total Ion Pairs	72	64	65	62	69
Intra A	31	26	29	29	25
Intra B	31	26	30	27	24
Inter AB	10	12	6	6	20

Figure 7.19b: Ion pairs, 5.0 Å cut-off.

	DS23R	Pig (closed)	*Ta*	*Ss*	*Pf*
Total Ion Pairs	98	82	60	82	105
Intra A	44	35	27	38	40
Intra B	44	35	23	35	39
Inter AB	10	12	10	9	26

Figure 7.19c: Ion pairs, 6.0 Å cut-off.

Looking simply at the total numbers of ion pairs, it can be seen that all the thermophilic CSs have a greater total number of ion pairs than the pig enzyme, but the psychrophile actually has the greatest total number of ion pairs. Looking then at the trends towards inter/intra subunit ion pairs, *Pf*CS has the most interface ion pairs but both *DS*CS and pigCS have more intersubunit interactions than the *Ta*CS and *Ss*CS and therefore the location of these ionic interactions is of more interest; are they stabilising otherwise disordered areas of the structure, or those which are particularly vulnerable to denaturation? (e.g. the termini, surface loop regions or the dimer interface).

Figure 7.20 lists the percentage involvement of particular residues in ionic interactions. These results (particularly the 4.0Å cut-off) show the considerably lower participation of charged residues in interactions in the pig enzyme when compared with any of the other CSs (in addition to having a lower overall charge content).

	DS23R	Pig (closed)	*Ta*	*Ss*	*Pf*
Total Arg/%	22/38=57.9	14/38=36.8	12/34=35.3	11/34=32.4	16/33=48.5
Total Lys/%	12/36=33.3	8/50=16.0	21/54=38.9	23/62=37.1	19/64=29.7
Total His/%	8/22=36.4	4/28=14.3	8/12=66.7	6/10=60.0	4/16=25.0
Total Glu/%	22/54=40.7	14/48=29.2	25/50=50.0	23/60=38.3	29/76=38.2
Total Asp/%	22/36=61.1	16/42=38.1	13/38=34.2	14/34=41.2	7/22=31.8

Figure 7.20a: Involvement of different residues in ion-pairs as a percentage based on total number of given residues in coordinate file and using a 4.0 Å cut-off.

	DS23R	Pig (closed)	*Ta*	*Ss*	*Pf*
Total Arg/%	28/38=73.7	22/38=57.9	26/34=76.5	26/34=76.5	31/33=93.9
Total Lys/%	18/36=50.0	22/50=44.0	35/54=64.8	35/62=56.5	40/64=62.5
Total His/%	18/22=81.8	18/28=64.3	8/12=66.7	6/10=60.0	11/16=68.7
Total Glu/%	36/54=66.7	34/48=70.8	34/50=68.0	40/60=66.7	55/76=72.4
Total Asp/%	24/36=66.7	26/42=61.9	24/38=63.2	19/34=55.9	17/22=77.3

Figure 7.20b: Involvement of different residues in ion-pairs using a 6.0 Å cut-off.

7.7.2 Ionic Networks

Using a 4Å cut-off, the multiple ionic interactions (three residues or more) were calculated due to the fact that these interactions are thought to be energetically more favourable than single ion pairs. The networks are summarised in figure 7.21 by means of the pairwise interactions involved, and the role they may play in stabilisation. There is not a vast difference in the total numbers or sizes of networks but the nature of these networks is different in the five CSs in the context of the dimer interface and loop regions.

7.8 Length of (and Ionic Interactions in) Loop Regions

It has previously been suggested that loop regions tend to be the most flexible regions of the structure, and are therefore often the first areas to be subject to proteolytic cleavage or heat denaturation [Dagget & Levitt 1993, Leszczynski, 1986]. It is possible that increased thermostability may be overcome by shortening loops or by additional interactions stabilising these regions. The equivalent loop regions of the five CSs have therefore been compared (several extra loops are present in the pig enzyme). The loop regions which have been compared can be seen in the schematic diagram with helices (C-S) labelled for a single *DS*CS monomer in figure 7.22.

	3 residues	Role	4 residues	Role	5 residues	Role
*DS*CS	D137-R140 D137-R141	Stabilises C-terminal end of helix I	R98-D204' K217-D204' D95-R98 D95-K217	Inter-subunit interaction stabilising each end of central helices G&M	—	—
	R150-D152 R150-E167	Stabilises loops I-J and J-K	E233-H230 H230-E226 H92-E226	Intra-subunit stabilisation of helix N at the dimer interface		
	K7-D15' K7-D359'	Interconnects N- termini of both monomers				
pigCS	R229-D208 R229-E226	Stabilisation of helices J and K	H246-E420 H246-E420' H246'-E240 H246'-E420	Single network stabilising dimer interface (helix L and loop S-T)	K7-E173 R117-E173 R117-D177 R117-E113	Interconnects helices E and I and K7 near the N-terminus
	E239-R421' E239-K423	Intersubunit (loops K-L and S-T)				
	R20- D428' R20-E17	Intersubunit connection of extended helix A with helix T				

	3 residues	Role	4 residues	Role	5 residues	Role
*Ta*CS	—	—	E188-H262 E188-R364' E188-H187	Interlinks loops O-P and K-L and forms inter-subunit interaction	—	—
			K74-E126 E126-R130 R130-E110	K74 in loops E-F interact with residues in helices G and I		
*Ss*CS	E169-R281 E169-R163	loop J-K interacts with loop P-Q	K14-E348 E348-H345 H345-D182	Stabilisation of helix S and interconnection with loop K-L and region of sheet near the N-terminus	K68-E71 K68-E72 K68-E118 E72-K126	Stabilisation of C-terminal end of helix E and interconnection with helix I
	D150-K154 K154-E164	loop I-J interacts with loop J-K				
	H183-E184 E184-H258	loop K-L interacts with loop O-P				
	E9-R259 E9-R355'	stabilises N-terminus at interface				

	3 residues	Role	4 residues	Role	5 residues	Role
*Pf*CS	—	—	E189-R356' E189-R358' R358-D12	Inter-subunit interaction stabilising dimer interface near the active site	H93-D113' D113'-K219 K219-D206' D206'-R99	Inter-subunit interaction stabilising each end of central helices G&M
			K276-E292 K276-E295 E292-E295 R296-E292	Stabilisation of helix Q and interconnection with helix P		

Figure 7.21: Ionic networks in the five citrate synthases ('denotes the residue is from the second subunit), displayed as the number of residues involved in a network, and the possible role the interaction may play in terms of stabilisation.

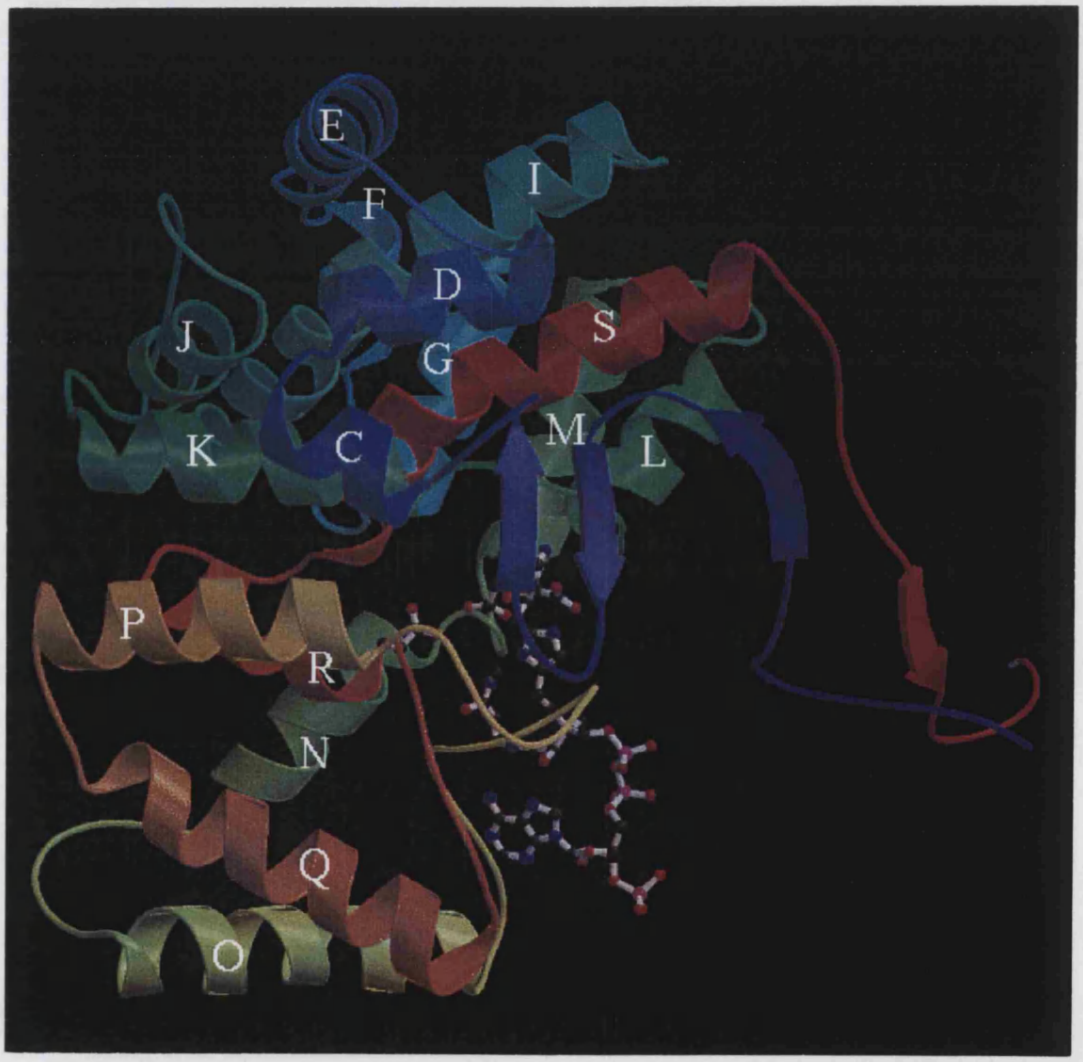


Figure 7.22: The *DS*CS citrate synthase monomer with labelling of helices C-R.

Loop	DS23R	pig	*Ta*	*Ss*	*Pf*
A-B		7			
B-C*		26			
C-D	0	9	2	0	0
D-E	3	3	3	3	3
E-F	4	3	5	4	4
F-G	6	4	5	4	4
G-I	7	3+helixH+5	8	7	7
I-J	11	12	11	11	11
J-K	3	2	3	3	3
K-L	4	6	4	4	4
L-M	3	3	2	2	3
M-N	4	5	5	5	4
N-O	12	4	3	3	3
O-P	17	15	18	18	14
P-Q	2	2	1	1	0
Q-R	4	9	6	10	11
R-S	5	6	5	4	6
S-T		10			

Figure 7.23: Table listing the number of residues in loop regions for the five CSs. *Loop B-C in the pig enzyme also contains a region of beta-sheet structure. Loop G-I in the pig enzyme also contains an additional helix.

Loop	DS23R	Pig	*Ta*	*Ss*	*Pf*
A-B					
B-C		D61-K325			
C-D					
D-E	E56-R375'			K55-E61	K57-E63
E-F	R74-E70		K74-E126 E126-R130 R130-E110	K68-E71 K68-E72 K68-E118 E72-K126	
F-G				E89-K108'	
G-I	H108-E121 R110-D113		K117-D118	K108-E89' K111-E114	H93-D113' D113'-K219 K219-D206' D206'-R99
I-J	E144-R139 R150-D152 R150-E167		E148-K151	R144-E147 D150-K154 K154-E164	K143-E327 K152-E169 E153-K73
J-K	R150-D152 R150-E167		K168-E158	D150-K154 K154-E164 E169-R281 E169-R163	E169-K152
K-L	R150-E167	E239-R421' E239-K423	D186-H351 E188-H262 E188-R364' E188-H187	H183-E184 E184-H258 K14-E348 E348-H345 H345-D182	E187-H344 E189-R356' E189-R358' R358-D12
L-M					
M-N					
N-O			D238-K291		E241-R293
O-P	D276-R278	K325-D61	K255-E235 E188-H262 E188-R364' E188-H187	H183-E184 E184-H258	R255-E236

Loop	DS23R	Pig	*Ta*	*Ss*	*Pf*
P-Q			K285-E172	E169-R281 E169-R163	
Q-R	K316-E310				
R-S					E327-K143
S-T		H246-E420 H246-E420' H246'-E240 H246'-E420			

Figure 7.24: Table listing ionic interactions in loop regions (including networks where at least one of the residues is situated in the loop region).

Loop C-D; PigCS has a loop containing 9 residues which is only 2 residues in *Ta*CS and absent in *Ss*CS, *Pf*CS and *DS*CS (Figure 7.25).

Loop D-E; All CSs have a 3 residue loop but a single ion pair is present in *DS*CS, *Ss*CS and *Pf*CS.

Loop E-F; The loop in PigCS is one residue shorter than *DS*CS, *Ss*CS, *Pf*CS and 2 residues shorter than *Ta*CS but the loops superimpose well and there are no ionic interactions in PigCS compared with a single ion pair in *DS*CS. *Ta*CS has a four residue intramolecular network and *Ss*CS has a five residue network which involves a residue at the N-terminal end of the loop.

Loop F-G; All loops are very similar but *Ss*CS has one residue involved in an inter-subunit ion pair with a residue in loop G-I in the other monomer (tying two loops together at the interface).

Loop G-I; The pig enzyme has a more extended loop containing the short helix H. This helix is absent in the other CSs. The other loops are very similar in topology to each other, with ion pairs present in *DS*CS and *Ta*CS. Both

*Ss*CS and *Pf* CS have inter subunit interactions (D113 in *Pf*CS being part of five residue inter-subunit network).

Loop I-J; All have similarly large loops but there are no ion pairs in pigCS with one ion pair in *Ta*CS. *DS*CS, *Ss*CS and *Pf*CS all have multiple ionic interactions linking loops I-J and J-K.

Loop J-K; The loop in pigCS is slightly shorter than the others containing no interactions, whilst the *Ta*CS loop has an ion pair. *DS*CS, *Ss*CS and *Pf*CS have interactions linking this loop with previous loop I-J.

Loop K-L; All loops are similar (the pig one being slightly longer). All CSs have ionic interactions stabilising this loop, with both *Pf*CS and pigCS having intersubunit interactions (*Pf*CS has 4 residue network).

Loop L-M; All loops well conserved, having similar topology and no ionic interactions.

Loop M-N; All loops well conserved, with no ionic interactions.

Loop N-O; This appears to be a long and flexible loop in *DS*CS which contains six charged residues (and no ion pairs). PigCS also has a longer and more extended loop than *Ss*CS and *Ta*CS (which both contain ion pairs) and *Pf*CS has the shortest loop.

Loop O-P; Although not obvious from the length of loops as designated by PROMOTIF, the loop in the pig enzyme is considerably more extended than the others. *DS*CS, pigCS and *Pf*CS all have single ion pairs stabilising this loop (interaction in the pigCS loop links it to loop B-C), with *Ta*CS and *Ss*CS loops having multiple ionic interactions which link loops O-P and K-L.

Loop P-Q; *Ta*CS and *Ss*CS loops both contain ionic interactions. In the case of *Ss*CS, this is in the form of a three residue network linking it with loop J-K. This loop is absent in *Pf*CS.

Loop Q-R; This loop is shortest in *DS*CS and it has already been suggested that the reason for this is that it seems to allow greater accessibility to the active site [Russell *et al*, 1998].

Loop R-S; There is no significant difference in the conformation of this loop in any of the CSs but *Pf*CS has an ion pair between this loop and loop I-J.

Although some of the differences in loop conformations (particularly near the active site) may be due to the open or closed nature of the structures, it is immediately obvious on comparison of the ionic interactions, that there is considerable difference between the five enzymes. There is a distinct absence of ionic interactions in the loops of the pig enzyme, with the thermophiles (and psycrophile) showing more extensive interactions. Many of these ionic interactions are involved with the dimer interface as well as those which interconnect one loop region with another, thus possibly 'pinning' both loops together. These results are summarised in figure 7.26. Although this is coupled with several examples of loop shortening (an example of which has been shown in figure 7.25) in the Archaeal when compared with the pigCS, there is not such an overall trend of this latter effect, as was first observed with initial pairwise comparisons [Russell *et al*, 1994].

Figure 7.25: Diagram showing loop C-D over-layed for the five citrate synthases (coloured DSCS; grey, pigCS; pink, *Ta*CS; yellow, *Ss*CS; orange, *Pf*CS; red).

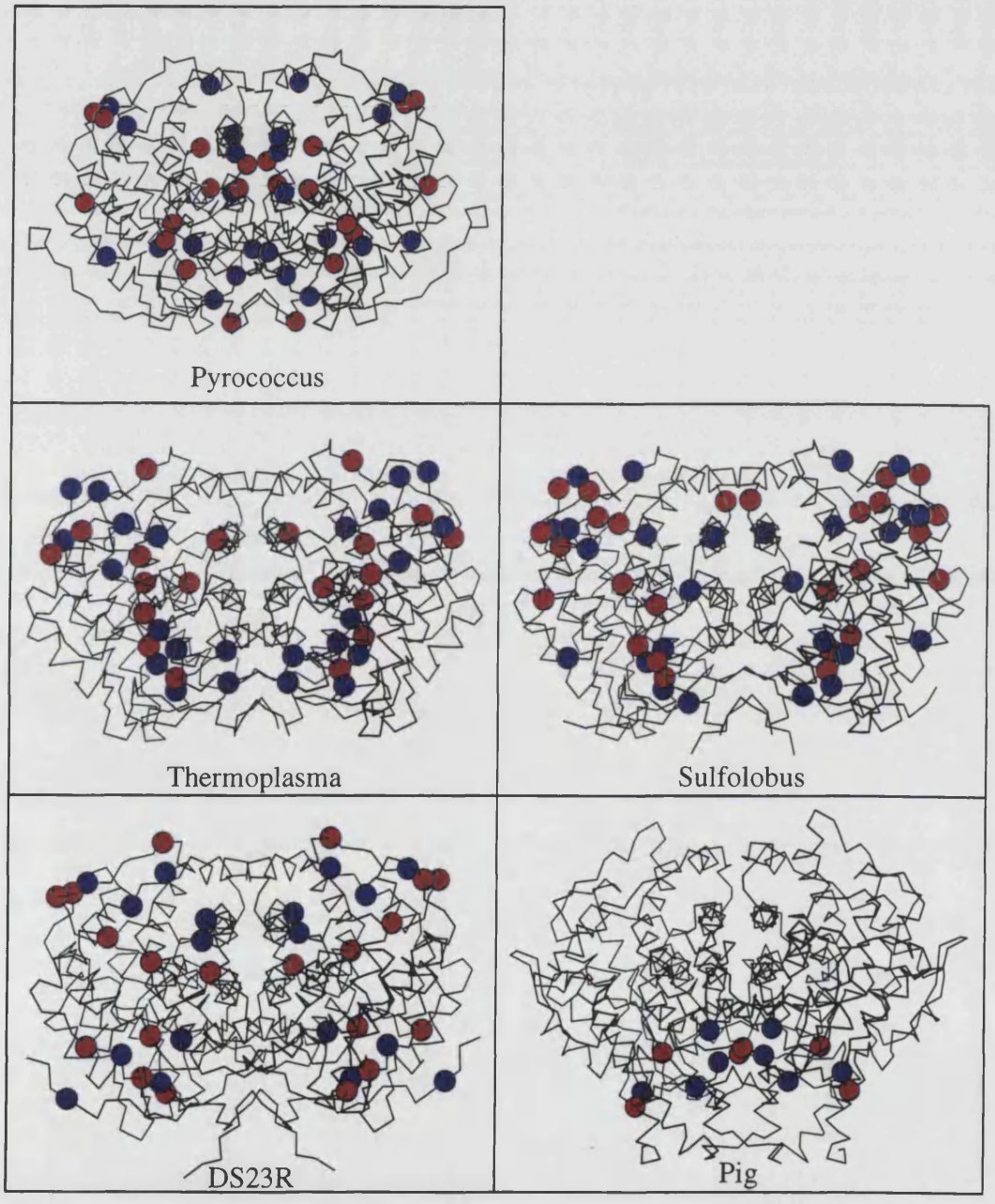


Figure 7.26: CPK sphere representation of all acidic (red) and basic (blue) amino-acids involved in ionic interactions where at least one of the residues is located in a loop region.

7.9 Dimer interface

The dimer interfaces were analysed according to several criteria as found in the table in figure 7.27 and ionic interactions were also analysed as before (figure 7.28). As *DS*CS, *Ss*CS and *Pf*CS have had the C-terminal arms removed prior to making these calculations, (to compare them all with *Ta*CS) the values generated do not represent the complete intersubunit contact. Complementarity can be viewed by means of a gap volume index [Jones and Thornton, 1995] defined as a gap volume between the monomers divided by ASA of interface. Better packing at the interface is therefore represented by a lower value for this gap volume index.

	DS23R	Pig (closed)	*Ta*	*Ss*	*Pf*
Interface ASA /Å^2	3403	4934	3154	3363	3698
% of total ASA	21.0	24.0	19.0	19.8	22.6
% polar atoms	34.8	37.8	32.6	32.3	39.2
% non-polar atoms	65.2	62.2	67.4	67.7	60.8
Hydrogen-bonds	44	42	24	28	54
Gap volume	10591	17164	6474	8474	9605
Gap volume index	1.52	1.74	1.03	1.26	1.29

Figure 7.27: Statistics calculated using the protein-protein interactions server [Jones & Thornton, 1995] for the PDB files with the C-terminal arm removed.

PigCS has a slightly different interface topology (particularly in the terminal regions) from the other CSs (as described earlier) and therefore has a considerably larger accessible surface than the other interfaces. It also has the highest percentage of the total ASA involved at the interface. *Ta*CS has the highest complementarity at the interface represented by the lowest gap volume (and lowest gap volume index) with *Ta*CS and *Ss*CS both exhibiting the highest percentage of hydrophobic residues compared with the other CSs. This coincides with both of these enzymes having the lowest number of hydrogen

bonds at the interface (with *Pf*CS having the highest number of hydrogen bonds). The three thermophiles all have lower gap volume indexes than both the pigCS and *DS*CS, suggesting better packing of the two subunits.

DS23R	**Pig**	***Ta***	***Ss***	***Pf***
K7-D15' K7-D359'	R25-D39'	E188-H262 E188-R364' E188-H187	E9-R259 E9-R355'	K8-D16'
R98-D204' K217-D204' D95-R98 D95-K217	D435-H28'	D205-K218'	D201-K214'	H93-D113' D113'-K219 K219-D206' D206'-R99
R375-E56'	E239-R421' E239-K423		E89-K108'	E189-R356' E189-R358' R358-D12
	R20-D428' R20-E17			R353-E11'
	H246-E420 H246-E420' H246'-E240 H246'-E420			R375-E48'

Figure 7.28: Ionic interactions at the dimer interface calculated with a 4Å cut-off.

Examining ionic interactions at the interface, the five CSs show considerable variation; interactions in the pig enzyme are all unique (with respect to the other CSs) and involve residues near the termini and outer helices of the eight-helical sandwich. The central helices of the dimer interface have no ionic interactions associated with them in marked contrast to the other CSs. Two ionic interactions are present between the last helix (helix T; not present in the other enzymes) and helix A (in the N-terminal extension) of the other subunit (D435-H28' and network involving E17', D428 & R20') and also a single ion pair between helix A of one subunit and helix B of the other (R25-D39'). One four residue network with two identical residues from each monomer (H246 in helix L

and E420 from loop S-T) and two three-residue networks linking E239 from loop K-L with R421 and K423 in loop S-T. The pig enzyme also has an intra subunit ion pair network associated with the N-terminus (R117, K7,E173, E113 & D177).

The other four citrate synthases all have interactions associated with the two internal helices of the interface (G&M) as well as those involved with the N-terminus and C-terminal arm. One completely conserved interaction is that between a totally conserved aspartate (D205 in *Ta*CS) at the N-terminal end of helix M with a lysine (at position 218 in *Ta*CS) at the C-terminal end of helix M in the other subunit. The lysine residue is not conserved in the pig enzyme. This leads to a single ion pair at each end of helix M in *Ta*CS. In *Ss*CS there is an additional ion pair in close proximity with the first one between E89 (loop F-G) of one monomer and K108 at the C-terminal end of helix G in the other, and thus both central helices have ion pairs at either end in *Ss*CS. In *DS*CS the first ion pair is part of a (3:1) four-residue network (in conjunction with D95 and R98 both in helix G). In *Pf*CS this is a five-residue (3:2) network with H93 and R99 (helix G) and D113 in loop G-I (giving two five-residue networks at this part of the interface). The four residue network in *DS*CS only comprises two single interactions directly across the interface, wheras the *Pf*CS five residue network has four such interactions (and five if calculations to 6Å are made). This would suggest that the *Pf*CS networks contribute considerably more to the inter-molecular interactions. The overall ionic interactions at the interface are summarised in figure 7.29 and more specifically those involved at the central helices G and M are shown in figure 7.30.

Examining the part of the dimer interface near the active site all the Archaeal and Bacterial citrate synthases have ionic interactions which also tend to stabilise the N-terminus, however *Pf*CS certainly has the most extensive ionic interactions with two four-residue networks (E189,D12,R356' & R358') and two single ion pairs (E11-R353'). Isoleucine clusters (see section 7.4) may also be indicative of strong hydrophobic interactions in this region in *Pf*CS and to a lesser extent in *Ss*CS and *Ta*CS.

Ionic interactions may also be important for prevention of fraying of the N and C termini and several of these interactions inter-link both terminal regions. The lengths of the C-terminal arms vary with *DS*CS being six shorter than *Pf*CS and *Ta*CS and five shorter than *Ss*CS. *DS*CS has less interaction of the C-terminal arm with the other monomer than the three thermophilic CSs (including one ion pair which appears to anchor the end of the arm; R375-E48' (*Pf*CS)). R375 and E48 are conserved in *Ta*CS and *Ss*CS suggesting the likelihood of this ion pair being present at the C-termini of the latter). *DS*CS also has an arginine residue (R375) which interacts with E56' but as this interaction is not directly at the end of the termini there may be more chance of fraying of this terminal arm in the psychrophile. Both N-termini in *Pf*CS also have an interconnecting ion pair (K8-D16') but this is a three residue interaction in *DS*CS (K7,D15' & D359'). *Ss*CS also has several terminal interactions (E9,R259 & R355') and *Ta*CS does only if interactions to 6Å are included. Including additional interactions to 6Å lead to a far more complex picture, particularly at the C-terminal arm region of *Ss*CS and *Pf*CS. There are also increased interactions near the C-terminus of *Ta*CS. In addition the two ion pairs at each end of helices G and M in *Ss*CS interact with each other. It is interesting to note that when including calculations to 6Å the pigCS does not have any additional ionic interactions involved with the dimer interface.

Figure 7.31 displays the buried face of the dimer interface across the four helices F,G,M & L and summarises the types of interactions taking place in this region. The mesophilic pigCS is shown to be predominantly a hydrophobic interaction (particularly helices F,L & M, with helix G containing several polar residues). In *Ta*CS there is an increase in hydrophobicity of helix G (the two serine and one threonine and residues lost with an increase from three to eight alanine residues in the thermophile). There is also the introduction of the charged groups at the ends of the central helices representing the single ion pair at each end of helix M. Much of this hydrophobicity is retained in *Ss*CS with the increase in charged residues at the ends of helix G & M. *Pf*CS shows the greatest charge content, involved in the two five-membered networks, with a less hydrophobic surface (in particular several polar groups in helix G). *DS*CS also displays a high degree of charge (corresponding to the two four-residue ion-pair networks) but a lower level of hydrophobicity.

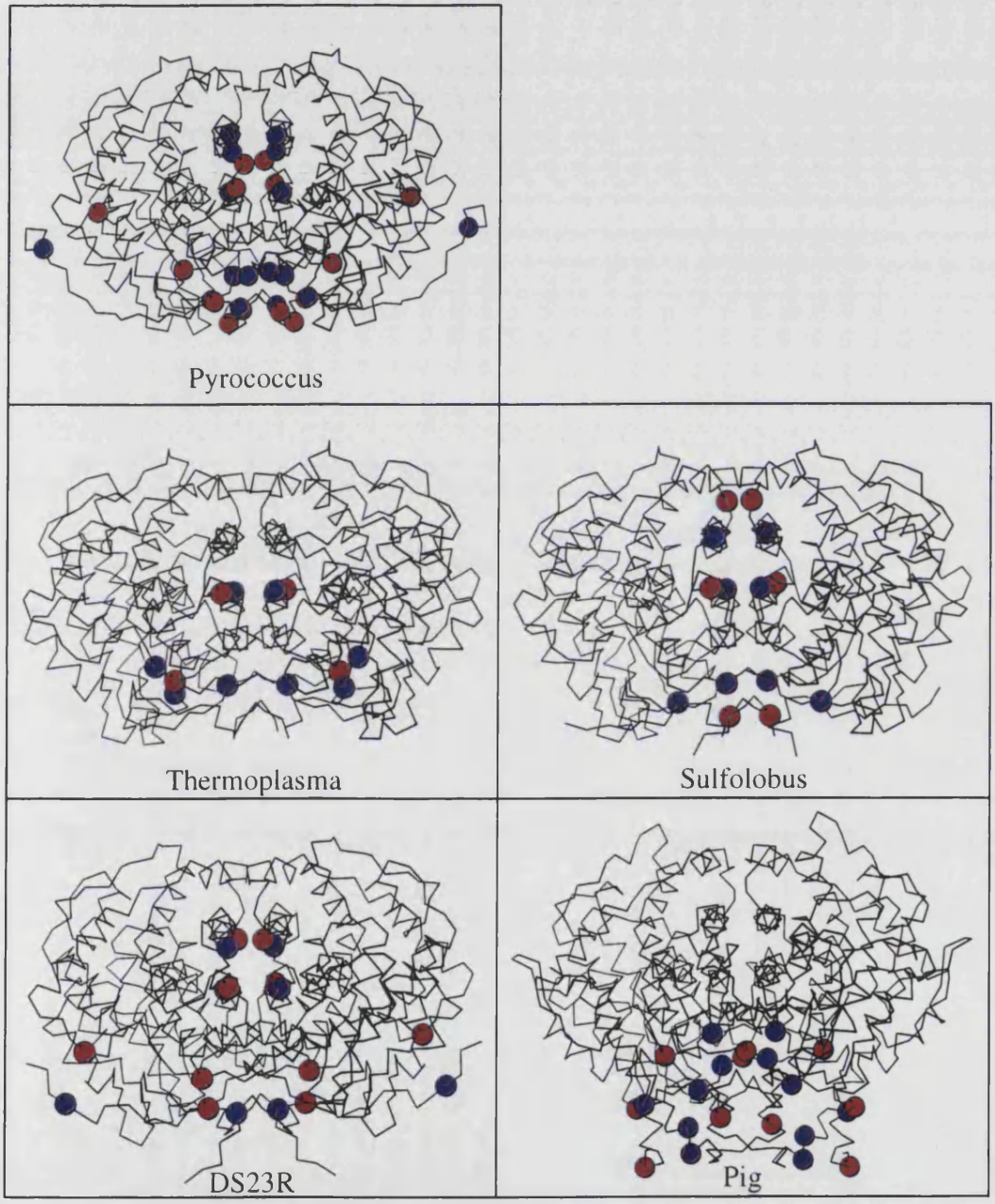


Figure 7.29: CPK sphere representation of all acidic (red) and basic (blue) sidechains involved with ionic interactions at the dimer interface.

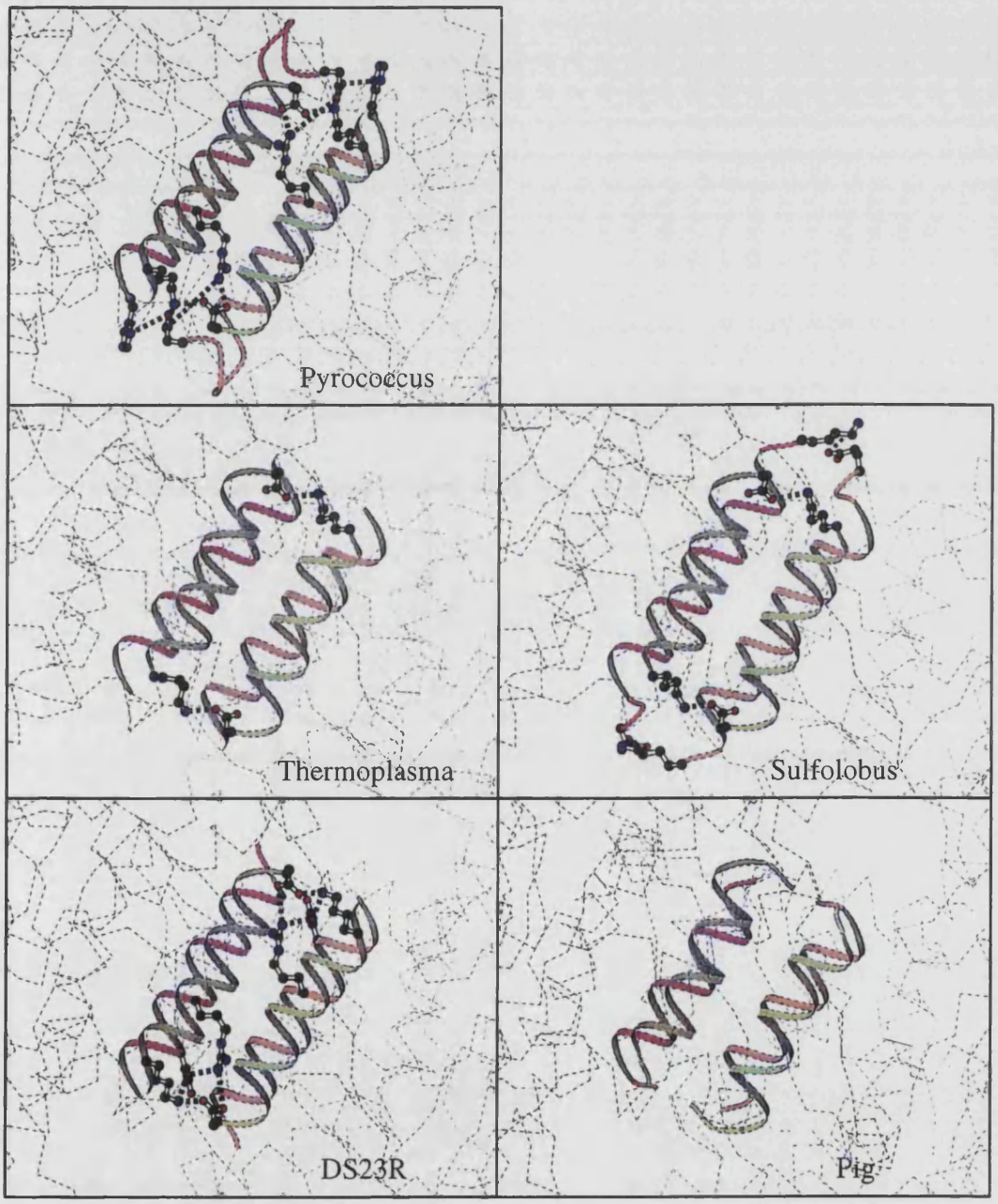
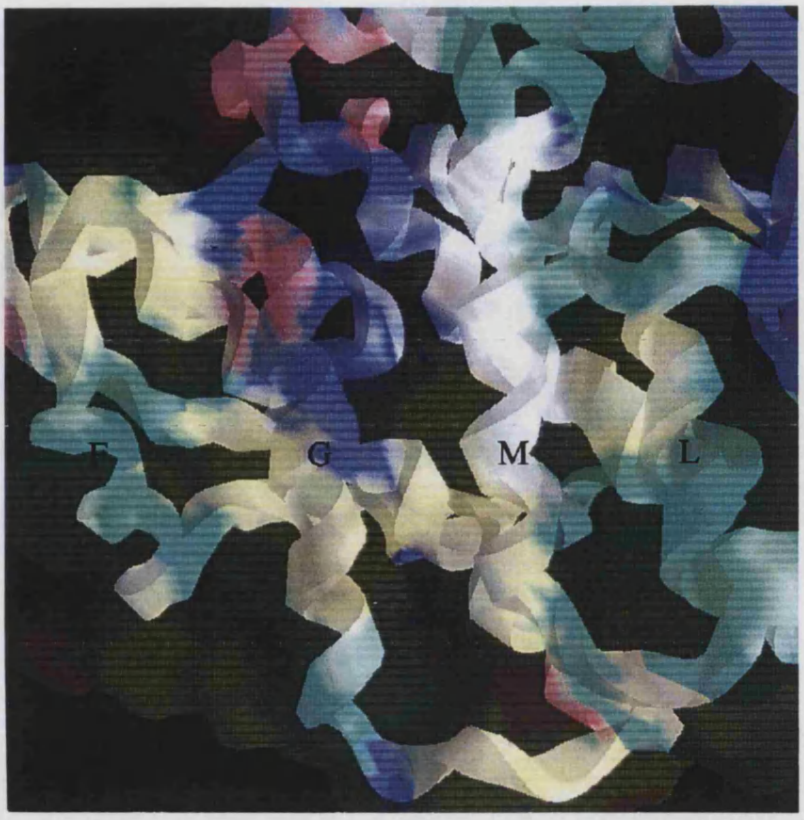


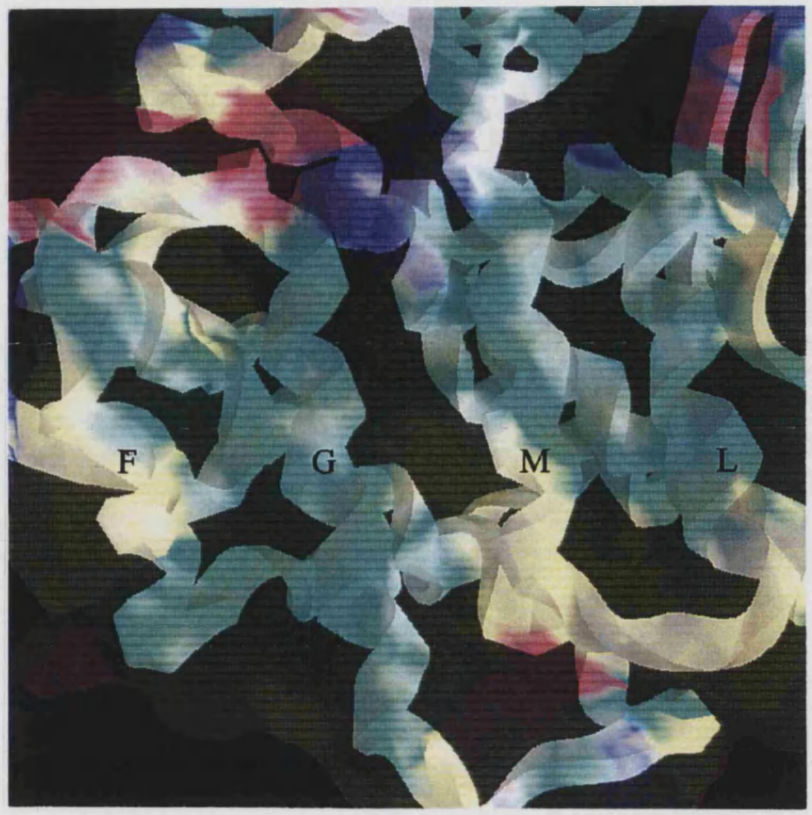
Figure 7.30: Diagram showing ionic interactions in the central helices (G green) and M (brown)) of the dimer interface.

a)

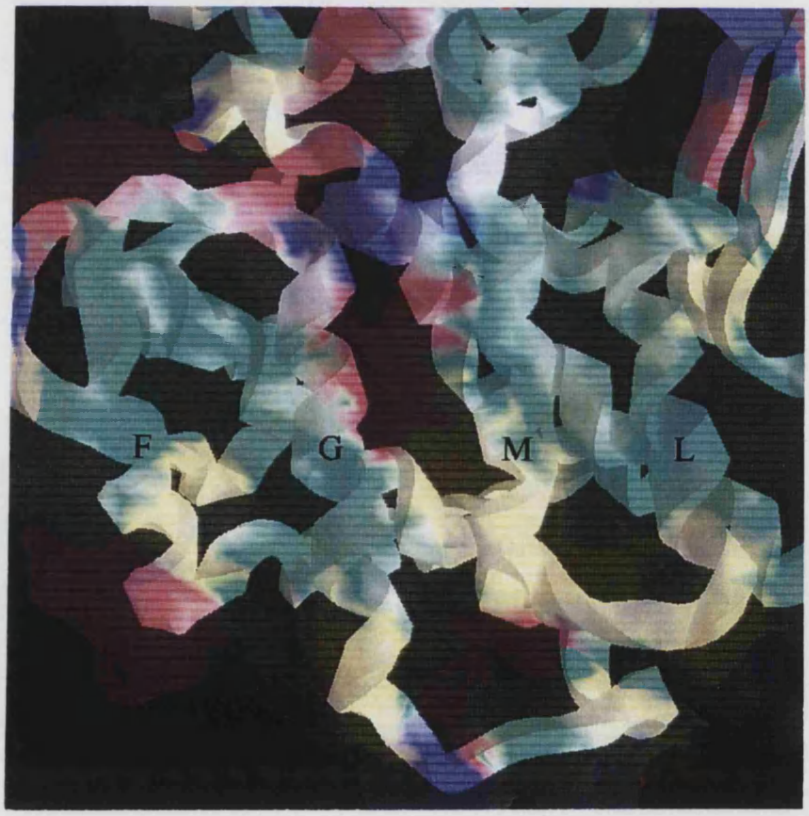


b)

c)



d)



e)

Figure 7.31: Buried face of the four interface helices F,G,M and L for a)*DS*CS, b)pigCS, c)*Ta*CS, d)*Ss*CS and e)*Pf*CS and is colour coded green; hydrophobic, yellow; polar, red; acidic (E,D) and blue; basic (R,K,H)). The diagram was created in **GRASP.**

7. 10 Discussion

Although not all structures (in both open and closed forms) are available to perform a complete analysis, the observations discussed in this chapter may correlate with the differing stabilities of these proteins.

Further insight into the heat inactivation of citrate synthase suggests that interactions at the dimer interface may be of great importance in the thermostability of citrate synthase. Data for the pigCS [McEvily & Harrison, 1986] shows that the dimer is in equilibrium with the monomeric form. Higher protein concentrations and the presence of substrates drive the equilibrium

towards the dimeric form. Inactivating the enzyme in the absence of substrates and at low protein concentration is a biphasic process due to the different rates of deactivation of the monomeric and dimeric forms; the slow dimer to monomer transition and rapid breakdown of the monomeric form. Data from Differential Scanning Calorimetry [H. Klump, unpublished data] has suggested that denaturation of *DS*CS is a two stage process; a reversible followed by an irreversible inactivation. An increase in enzyme activity preceding the initial inactivation is observed (and can be repeated on cooling and subsequent re-heating of the enzyme). Initial ultracentrifugation experiments have shown that this first stage is likely to be a dimer to monomer transition before unfolding of the monomers in the second irreversible stage. A single peak is observed on examination of the thermostable citrate synthases but asymmetry of this peak suggests that it is possible for this to be a small energy contribution followed by a larger one. This could represent a similar process to that found for *DS*CS, with an immediate unfolding following the initial dimer dissociation at higher temperatures.

Overall 3-D structures;

In general, as with so many other thermophilic proteins, the 3-D structures are very similar with largest deviations in the large domain possibly indicating that differences in thermostability are due to changes at the dimer interface. A domain-swap mutant [M. Arnott, unpublished data] comprising *Pf*CS large domain and *Ta*CS small domain has a thermostability only slightly lower than that of the wild type *Pf*CS. Similarly, a mutant enzyme with *Ta*CS large domain and *Pf*CS small domain has a thermostability considerably lower than that of the wild type *Ta*CS. These results do not discount the former supposition.

Amino acid content;

The thermophiles exhibit an increase in charged residues (although a reduction in histidine which is relatively thermolabile) with concomitant reduction in polar and uncharged residues when compared with the pig enzyme whilst *DS*CS also has a reasonably high charge content. Glutamate is perhaps also favoured over aspartate due to the formers greater helical propensity and

lower likelihood of chain cleavage at high temperatures [Daniel *et al*, 1996]. There is also a tendency towards fewer thermolabile residues; cysteine, methionine in the thermophilic enzymes compared with both pigCS and *DS*CS. The thermophiles have a lower glutamine content than the pigCS enzyme. *Pf*CS also has a lower content of asparagine. *DS*CS however, also has lower contents of cysteine and glutamine than the pig enzyme. Of the hydrophobic residues, isoleucine seems to be favoured in the thermophiles but the overall hydrophobic content shows little difference. The only aromatic residue showing a consistent trend is tyrosine which increases in content up the temperature scale (perhaps due to its additional hydrogen-bonding potential). Analysis of glycine and proline residues shows that there is no significant trend in the total contents of each amino acid, although perhaps location in the structure is more important; the thermophiles have a lower number of intra-helical prolines than both pigCS and *DS*CS (which has been shown to be destabilising as discussed in the introduction). The psychrophile also has fewer prolines in turn and loop regions than the other CSs and the pigCS shows the highest content of glycines in the loop/turn regions, both observations serving to increase the entropy of the unfolded state (with respect to the thermophiles) by increasing chain flexibility.

Stability of secondary structure (α-helices);

Initial analysis of the helical regions in terms of the presence of β-branched residues, i(i+3) and i(i+4) ion pairs and helix dipole interactions show that there does not appear to be a significant difference in the stability of helices of the thermophiles and mesophiles (perhaps the most obvious trend being a lower number of proline residues at an internal position in the thermophiles). A more sophisticated analysis could however be carried out to quantify these factors more thoroughly.

Compactness;

As discussed in the introduction, compactness and rigidity of heat stable proteins has often been found to be synonymous with their thermostability. Rigidifying features such as the presence of prolines or absence of glycine

residues has been mentioned above. The compactness can be described in a number of other ways. There is a tendency towards smaller ASA and volume when comparing the Archaeal CSs with the pigCS. Although the total percentage of atoms buried is similar for all the CSs, the decreased burial of hydrophobic groups of *DS*CS compared with the other CSs (and which is highest for *Pf*CS) perhaps exhibits the lower requirement for hydrophobic stabilisation at psycrophilic temperatures (below). There may be more efficient packing in the protein core and although aromatic interactions have not been studied, the hydrophobic isoleucine clusters are certainly present in the thermophiles. This may also be indicative of better Van der Waals interactions and tighter packing particularly at the interface. The increased greater dimer interface complementarity (measured by the gap volume index) in the thermophiles may also be a significant feature. The tendency towards fewer cavities in the thermophiles should also correlate with the improved packing of these proteins.

Ionic interactions;

The Archaeal and Bacterial CSs have a higher total number of ionic interactions (calculated with the typical 4.0Å cut-off) than the pigCS which also exhibits the lowest percentage participation of charged residues in ion pairs or networks. *DS*CS actually has the most ionic interactions (perhaps for the reason discussed below). *Pf*CS has the most extensive interactions across the dimer interface whilst *DS*CS has more than both *Ta*CS and *Ss*CS. PigCS has a higher number of inter-molecular ionic interactions than the thermophiles but these are all unique with respect to the others. It is also interesting to note that calculating all possible interactions to 6.0Å, the number of inter-molecular interactions only increase for the thermophilic proteins. The total number of ionic networks (which should be favoured for the entropy argument) is not significantly different between the proteins but there are different patterns in the key areas thought to be crucial for stabilisation; loops, termini and particularly the dimer interface.

Loop regions;

There is a tendency towards shorter (even absent) loop regions in the thermophiles correlating with the compactness of these proteins, when compared with pigCS. Many of these shorter loops are similarly short however in *D*SCS (apart from loop N-O) . A more dramatic difference is seen in the comparison of the ionic interactions in these regions; with very few being present in pigCS and a large number occurring in the Archaeal and Bacterial proteins, and the thermophilic CSs (particularly *Ss*CS) having the the most extensive networks which cross-link loop regions. This compares with the observations of β-Glycosidase from *Sulfolobus solfataricus* [Aguilar *et al,* 1997] which was shown to have ionic interactions (specifically networks) over the surface of the protein such that they cross-linked areas of surface structure. This may also be the case in citrate synthase, where ionic interactions play a role in stabilisation of these otherwise more disordered parts of the protein.

Dimer interface and Termini;

The eight-helical sandwich part of the dimer interface shows a definite trend towards increasing hydrophobicity going from *D*SCS and pigCS to *Ta*CS and *Ss*CS. *Pf*CS also has a greater degree of hydrophobicity in this region than *D*SCS and pigCS but lower than the other two thermophiles. The overall complementarity of the dimer interface also seems to be greater for all of the thermophiles (specifically *Ta*CS) than either *D*SCS or pigCS. The ionic interactions in the central helices (G and M) also show the increase from none in pigCS, two single ion-pairs in *Ta*CS, four single ion-pairs in *Ss*CS (which would be classed as a network using a 6Å cut-off) and the two five-residue networks in *Pf*CS. *D*SCS also has the two four-residue networks here but these seem to be less extensive than those in *Pf*CS (with fewer interactions actually across the interface). SDM experiments have been carried out to analyse the importance of the dimer interface networks in *Pf*CS [M. Arnott, unpublished data]. A D113A mutant disrupting this five-residue ionic network resulted in a considerable drop in thermostability with the introduction of the hydrophobic side-chain (to approximately one third of the wild type *Pf*CS). A D113S mutant was slightly more stable than the former (although still much less stable than wild-type *Pf*CS) probably due to the hydrogen-bonding

potential of the serine hydroxyl group. *Pf*CS also has the additional two four-residue networks near the active site region (with the other Archaeal and Bacterial enzymes displaying interactions to a lesser degree) as well as the isoleucine cluster which is partly conserved in *Ta*CS and *Ss*CS. The latter feature may be favoured in the thermophiles due to an optimisation of hydrophobic interactions in the protein core, but this cannot be confirmed without a more thorough analysis of atomic packing. All the ionic interactions in the pig enzyme are associated with one part of the interface (and not the helical sandwich) and thus have no direct comparison with interactions in the other enzymes. The parts of the dimer interface associated with both termini seem to be stabilised by ionic interactions particularly in the *Pf*CS, but also to some degree in *DS*CS and *Ss*CS. Differences in the C-terminal arm (which is not present in pigCS) of *DS*CS with respect to the thermophiles are reviewed below. In the three thermophiles (the *Ss*CS and *Ta*CS were not seen in the electron density maps) the end of the C-termini seem to have a single ion pair 'anchoring' the end of the arm. Again, SDM studies on *Pf*CS have shown the importance of these interactions; a -2 residue mutant (removing the first ion pair) has a thermostability which is approximately a third of the wild-type *Pf*CS, whilst removing the whole of the C-terminal arm (-13 residue mutant) gives a stability in the region of two thirds that of the wild type [Michael Arnott, unpublished data]. This can be rationalised such that removal of a small part of the terminus may lead to a flexible C-terminus which could begin to 'unwrap' the two monomers, wheras removal of the rest of the arm prevents this flexible region from initiating unfolding. These results compare well with other studies [Meinnel *et al*, 1996].

***DS*CS;**

Although there are several trends discussed above which are evident on comparison of pigCS with the three thermophilic enzymes, *DS*CS exhibits many of the features that are being associated to protein thermostability. The psychrophile has the highest total number of intra-molecular ion pairs (and also many ionic interactions stabilising loop regions and at the dimer interface). This increase in ionic interactions (and indeed decreased burial of hydrophobic surface) may correlate with a decrease in the hydrophobic effect at lower temperatures and a need for additional stabilisation (to prevent cold

denaturation). This increased hydrophobic exposure should also reduce stability at higher temperatures. Also, the presence of only one or two weak parts within the structure may be enough to initiate unfolding at higher temperatures. In particular, some of the features which have been proposed to be responsible for the cold activity [Russell *et al*, 1998] of this enzyme may contribute towards its heat lability (a small domain with high flexibility relative to the large domain and the presence of a very flexible charged loop; N-O). If the first stage of denaturation is in fact dimer dissociation (as suggested by the Differential Scanning Calorimetry) then perhaps subtle differences at the interface such as the lack of such strong networks in the central helices G&M and overall less ionic interactions accross the interface when compared with the *Pf*CS without the extent of hydrophobic interaction that are present in the *Ta*CS or *Ss*CS (especially in the region of the eight-helical sandwich). The lower contact of the C-termini with the second monomer in *DS*CS may also be important; although there is an ion pair which may be anchoring the end of the C-terminal arm, this arm is 5-6 residues shorter than the thermophilic CSs. There are also three residues following the ion pair stabilising the end of this arm in *DS*CS and this may result in the terminus being more susceptible to fraying. SDM studies on this part of the structure in *Pf*CS have been shown to have large effects on the thermostabilty (above).

Summary of stabilising features found in thermophilic CSs;

In general it seems that in the case of citrate synthase, the already structurally sound areas such as the α-helices seem to show few obvious differences between the species in the initial crude analysis and this is similar to studies previously carried out on thermophilic helices [Facchiano *et al*, 1998]. The regions of the structures which are either otherwise structurally weak and disordered or are obvious initiation points for unfolding show the greatest deviation. In order for a protein to be heat stable, it must exhibit stabilising features, but alternatively, the presence of only a small number of labile regions could result in a vast difference in overall thermostability (as has been shown by simple SDM studies eliminating only one or two interactions). It is therefore these specific areas which are in greatest need of stabilisation, and in particular, the dimer interface and termini, as well as loop regions between areas of secondary structure are all likely to be key areas. The general trend

towards more efficient packing and burial of hydrophobic residues, rigidifying flexible regions (by either mutations of the form $X_{aa} \rightarrow$Pro and Gly$\rightarrow X_{aa}$, shortening loops, or the introduction of ionic interactions such that they crosslink areas of surface structure) and reduction of thermolabile residues all seem to be important at some level. Focusing then on the dimer interface, the difference in hydrophobicity and packing, coupled with the introduction of ionic interactions (particularly networks) may account for most of the differences in stabilities of the five enzymes. These findings correlate well with the growing number of studies which conclude that ionic interactions stabilising crucial areas of structure are perhaps the most singular method of stabilisation of proteins at hyperthermophilic temperatures.

CHAPTER 8

Final Discussion and Future Work

Elucidation of the crystal structure of *S. solfataricus* citrate synthase (which has a stability in between that of the thermophilic *T. acidophilum*, and hyperthermophilic *P. furiosus* enzymes) has allowed a complete comparison of the thermostability of this enzyme family, as summarised in the previous chapter. This has added to the work already carried out on these enzymes [Russell *et al*, 1994, 1997, 1998] and amongst other factors, reinforced the importance of interactions at the dimer interface. Again results point to the improvement of hydrophobic interactions particularly in the thermophiles, and of optimised electrostatic interactions, both at very high and indeed very low temperatures. These observations can be explained by a simplistic view of the theoretical arguments as introduced in chapter one. The increase in contribution from hydrophobic interactions to protein folding with increasing temperature is likely to lessen at very high temperatures (with the entropic contribution to the hydrophobic free energy tending towards zero at approximately 115°C [Baldwin, 1986]). There is also a reduced significance of hydrophobic free energy at low temperatures. These effects are due to a reduction of the entropic gain on desolvation of hydrophobic side-chains; at very high temperatures the ordering of water molecules into a 'clathrate' structure may be lessened due to their very high mobility and conversely there will be a very low mobility of water molecules at psychrophilic temperatures [Russell *et al*, 1998]. Although, at 100°C, the hydrophobic effect is likely to remain significant, [Dill *et al*, 1989] the high conformational entropy of the protein chain (when compared with mesophilic temperatures) may also be counterbalanced by chain rigidifying, shortening, or, as displayed most strikingly in this case, by the introduction of electrostatic interactions at denaturation hot-spots. The additional argument favouring electrostatic interactions at high temperatures is the lower desolvation penalty for the formation of a salt bridge [Elcock *et al*, 1998].

Although several Site-Directed Mutants have been created of both the *Ta*CS and *Pf*CS enzymes to probe for the most significant thermostabilising features (see discussion in previous chapter) it is hoped to expand on this work. The determination of both the open and closed structures from all five organisms would also allow an unequivocal study to be made, eliminating any suppositions which may be due to the presence of substrates in the active site.

Comparison of the citrate synthase study with other examples where several thermophilic enzyme structures have been determined (i.e. not simply a pairwise comparison) show a degree of similarity on a broad scale but are a reminder that there may be important specific factors in each case. The findings correlate well with that of glutamate dehydrogenase, emphasising the importance of interface interactions showing an increase in hydrophobicity and introduction of ionic interactions for the *T. maritima* GLUDH interface, with a move towards more extensive ionic stabilisation in these regions in *P. furiosus* [Knapp *et al,* 1997]. Also, the hyperthermophilic *T. maritima* GAPDH has an increased total number of ion pairs and of intra-molecular ion pairs compared with mesophile and thermophile, [Korndorfer *et al*, 1995] however the *B. stearothermophilus* enzyme (moderate thermophile) actually has more inter-subunit ionic interactions than the hyperthermophile. Although highly specific in terms of the structural context, similarities demonstrate that general mechanisms can still be drawn from these studies.

Unfortunately it is not possible to analyse the structure of *Pw*TIM as part of this project, however the continuation of this work may soon enable the comparison with TIMs from organisms growing at a similarly wide range of temperatures; the thermophilic *B. stearothermophilus*, hyperthermophilic *T. maritima* and the psychrophilic *V. marinus*, in addition to the mesophilic structures.

Although the *Pw*TIM tetramer was considered too large (96kDa) for analysis by the Nuclear Magnetic Resonance spectroscopy (NMR), a preliminary NMR experiment was carried out to assess the viability of such a study. The advantage of working with a thermostable protein is that the NMR experiment

can be carried out at higher temperatures whilst retaining structural integrity of the enzyme. Several 1-D spectra were measured using a 500MHz NMR operating at three different temperatures (30°C,60°C and 80°C). These spectra are shown in Figure 8.1. This work was carried out by Prof. Steve Homans at the University of St. Andrews. It can be seen from the spectra that greater definition is observed in the 80°C spectra when compared to the other operating temperatures as linewidth decreases with increasing tumbling times. This is most prominent in particular areas of the spectra such as at a chemical shift of 0.5-1ppm corresponding to the aliphatic side-chain protons and that corresponding to the main-chain N-H and aromatic side-chain protons (chemical shift 7-8ppm). Such a large molecule would require hetero-nuclear methods (labelling by ^{13}C and ^{15}N) for the assignment of all protein resonances, and it is perhaps not feasible to carry out full analysis on *Pw*TIM. However this highlights the possibility of using NMR to study these enzymes at temperatures more similar to those at which they naturally operate, and add to the information which we already know from analysis of crystal structures.

As has been mentioned above, experimental back-up is needed to couple crystal structure observations with Side-Directed Mutagenisis, by removing the proposed stabilising factors in these thermophilic enzymes. The problem here is the complexed nature of these systems, in which the interactions are heavily correlated. This creates the possibility of introducing subtle 'knock-on' effects as the result of a given mutation which can obscure results. In addition, although knowledge may be gleaned by the successful removal of a stabilising interaction, the ultimate aim; engineering thermostability, is likely to remain a difficult problem. There are relatively few examples where a substantial degree of stability has been successfully engineered into mesophilic proteins whilst retaining catalytic efficiency. However, a particularly striking example is that of the thermolysin-like protease from *Bacillus stearothermophilus* [Van den Burg *et al*, 1998] where a considerable increase in thermostability was achieved. In this case an eight-fold mutation, mainly increasing the rigidity of a single flexible surface loop region, in addition to the introduction of both a salt bridge and disulfide bridge, created a mutant enzyme stable at 100°C. Remarkably, this enzyme had similar activity to the wild-type enzyme at 37°C.

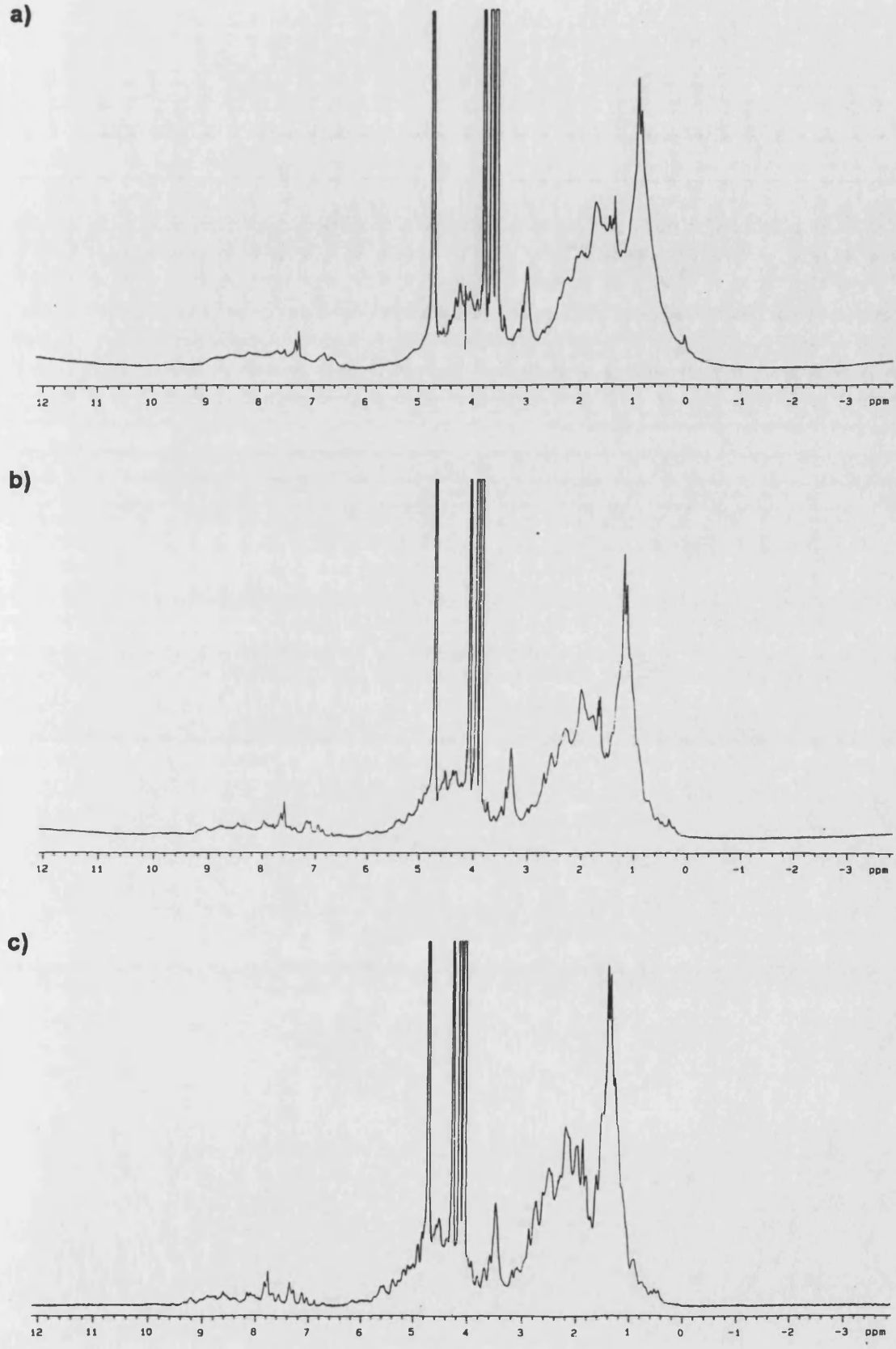


Figure 8.1 : 1-D NMR spectra measured on the PwTIM at a)30°C b)60°C and c)80°C.

The development of algorithms for predicting and or designing mutant enzymes with increased stability [Lee & Levitt, 1991, Viguera *et al*, 1996, Malakauskus & Mayo, 1998] have thus far been limited in their use, and the general application to different systems has yet to be demonstrated. Most studies have therefore relied simply on specific structural observations or comparison with mesophilic enzymes to identify stabilising interactions or denaturation hot-spots. In several cases, a cumulative effect has been observed by concurrent point mutations, [Zhang *et al*, 1992, Kimura *et al*, 1992, Akasko *et al*, 1995, Bogin *et al*, 1998] thus favouring a model where global unfolding processes are rate limiting. Other examples of isolated mutations show non-additive effects suggesting the importance of local unfolding processes [Hardy *et al*, 1994]. Perhaps much can still be learned from the screening of Random Mutagenesis experiments [Mathews, 1987, Alber & Mathews, 1987] due to the possibility of overlooking silent mutations (in which the initial mutation creates a mutant enzyme less (or equally) stable than the wild type but subsequent mutations can then produce a more stable enzyme) as it has been shown that even when working with highly homologous models, it is difficult to conclude the reasons for such effects without structural information [Vetriani *et al*, 1998].

Whilst attempting to ascertain whether or not it is possible to achieve enzyme stability at even higher temperatures, it should also be noted that factors such as the stability of metabolites -and not in fact enzyme stability itself- may limit the upper growth temperatures of life [Danson *et al*, 1998]. Focusing again on the ultimate aims of this work; being to fully understand the nature of these structure-stability relationships in the hope that stability can be engineered into a protein of choice, the practical answer in terms of biotechnological application, is probably to isolate the desired enzyme from an organism in the appropriate temperature range. This option is becoming far more feasable with the increasing number of complete Genome Sequencing Projects, [Deckert *et al*, 1998] and the hidden potential of the many novel enzymes which these Archaea contain, may provide the richest resources of all [Danson *et al*, 1998].

APPENDIX 1

List of Programs

ALSCRIPT	Displays multiple sequence alignments
AMORE	Molecular Replacement package
BOBSCRIPT	Extension to MOLSCRIPT
CAD	Manipulation of reflection data
COMPOSER	Protein structure analysis/alignment package
CONTACT	Calculates protein structure contacts
DENZO/SCALEPACK	Data processing/scaling
DM	Density modification
DSSP	Calculates protein secondary structure from coordinates
ECALC	Calculates normalised structure factor amplitudes
ENVIRONMENTS	Assesses environment of amino-acid residues in protein structure
EXTEND	Extends electron density maps over required volume of unit cell
FFT	Calculates fast Fourier transformation
FHSCAL	Uses Kraut scaling procedure to scale derivative to native data sets
FREERFLAG	Assigns FREER flags to reflections
GCG	Sequence analysis package
GRASP	Displays protein structure properties
HBPLUS	Calculates hydrogen-bonds in protein structures
HKLVIEW	Displays zones of reciprocal space
LSQKAB	Fits atomic coordinates
MAPMAN	Manipulation of electron density maps
MLPHARE	Phase calculation and refinement
MOLSCRIPT	Schematic representations of protein structure
MULTALIGN	Performs multiple sequence alignments

MTZMNF	Replaces missing data in MTZ file with missing number flags
MTZUTILS	Manipulates reflection data files
NCSMASK	Creates non-crystallographic symmetry masks
NPO	Plots Patterson maps and structures
O	General model building/ manipulation
OOPS	Real space refinement
PDBSET	Manipulates PDB files
PEAKMAX	Searches for peaks in maps
POLARRFN	Performs Self rotation funtion (spherical polar coordinates)
PROCHECK	Protein structure validation
PROMOTIF	Calculates protein secondary structure from coordinates
PROTIN	Prepares restraints for REFMAC
REFMAC	Refine protein structures using least squares or -loglikelihood residuals
REPLACE	Molecular replacement package incorporating locked rotation function
RFCORR	Analyses correlation between cross and self rotation functions
SCALEIT	Performs derivative to native data scaling
SFALL	Calculates structure factors using FFT
SHELX	Heavy atom site searches
SIGMAA	Phase combination
TFFC	Performs translation function
TRUNCATE	Converts reflection intensities to amplitudes
UNIQUE	Generates unique reflection data set
VECREF	Vector space refinement of heavy atom sites
VECSUM	Difference Patterson solution
VOIDOO	Analyses cavities in protein structures
XPLOR	General protein crystallography and NMR package including structure refinement and molecular replacement

APPENDIX 2

Crystal Symmetry

A crystal may be built by regular assembly of a basic building block, known as the unit cell. The unit cell can be defined by a parallelipiped with axial lengths a,b,c and angles between these axes being α, β, γ (figure 1). The points at the corners of the unit cell are called the space lattice. The asymmetric unit is the simplest object within the unit cell, which can be related to other identical objects by a set of symmetry operations particular to the space group of the unit cell.

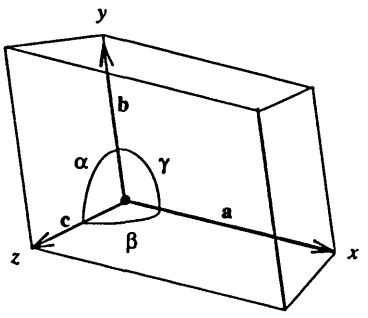
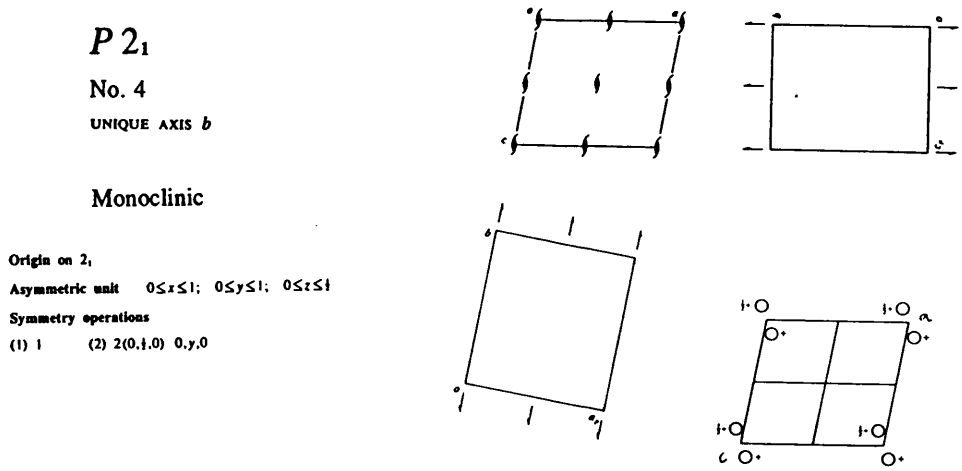


Figure 1: the unit cell

There are thirty two crystal classes (or point groups) defined by the external symmetry elements present. However, proteins can form crystals of point groups which contain rotational symmetry only, due to their enatiomorphic nature. Also, seven crystal systems (triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal and cubic) exist depending on the equivalences of the unit cell dimensions described above. The nature of the space lattice can vary for these crystal systems, giving rise to fourteen possible Bravais lattices. The additional presence of internal translational symmetry elements (screw axes) create the sixty-five possible space groups which a protein crystal may exhibit.

Figure 2 shows the location of symmetry axes, and symmetry equivalent positions for the monoclinic $P2_1$ and orthorhombic $P2_12_12$ space groups (in which *Pw*TIM and *Ss*CS were found to crystallise). In space group $P2_1$ (with unique axis b), there are equivalent positions at (x, y, z) and (-x, y+1/2 ,-z). For $P2_12_12$ this becomes (x, y, z), (-x, -y, z), (-x+1/2, y+1/2, -z) and (x+1/2, -y+1/2, -z). The rotational symmetry of the unit cell creates symmetry within the diffraction pattern. There are also reflection conditions brought about by the presence of screw axes (for example in the space group $P2_1$, systematic absences occurr for 0k0 reflections when k is even).



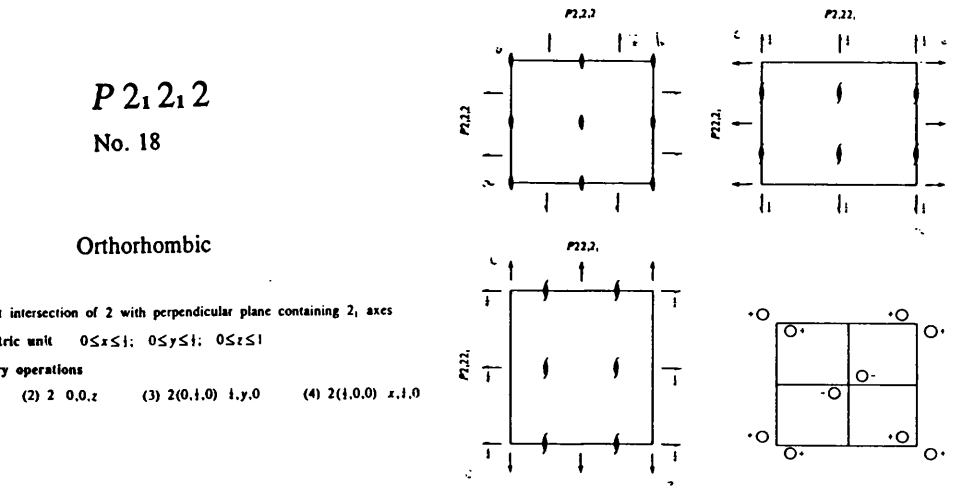


Figure 2: Symmetry equivalent positions for $P2_1$ and $P2_12_12$ unit cells (from International tables for X-ray crystallography Vol. 1, International Union of Crystallography).

Miller Indices hkl (where h, k, and l are integers) are used to define a plane throughout the crystal which cuts the unit cell at a/h, b/k and c/l. The lattice of the unit cell is in real space and related to the diffraction pattern by a reciprocal lattice (a*,b*,c*) with dimensions inversely proportional to the dimensions of the real lattice. In reciprocal space, h, k, and l, are indices of a point in the reciprocal lattice and index individual reflections of the diffraction pattern.

Diffraction from a Crystal

i) Braggs Law (Real Space)

Bragg first demonstrated the ωndition for diffraction from two parallel planes in a crystal lattice [Bragg, 1913]. A diffracted beam is produced from two incident beams R_1 and R_2 only when θ meets the following condition.

$$2d\sin\theta = n\lambda$$

where d is the interplanar spacing, θ is the angle of incidence (and angle of reflection), n is the pathwave difference integer and λ is the wavelength of the incident beam

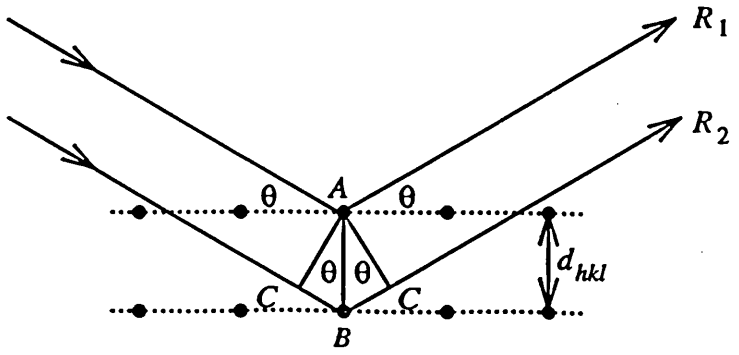


Figure 3: Braggs law (BC = dsinθ) [Rhodes, 1993].

ii) Reciprocal Space

The geometrical construction described by Ewald [Ewald, 1921] can be used to describe Braggs Law in reciprocal space. A sphere of radius $1/\lambda$ can be drawn with the centre at a point C, and as the crystal is rotated about the origin point O, the reciprocal lattice points cross the surface of the Ewald sphere (intersecting at points P and P'), fulfilling the Bragg condition, and thus incident beam X gives rise to diffracted beam R (figure 4).

As the angle PBO is θ

$$\sin\theta = OP/BO = OP/(2/\lambda)$$

$$\Rightarrow \quad 2d(hkl)\sin\theta = n\lambda \qquad (\text{as } OP = 1/d(hkl))$$

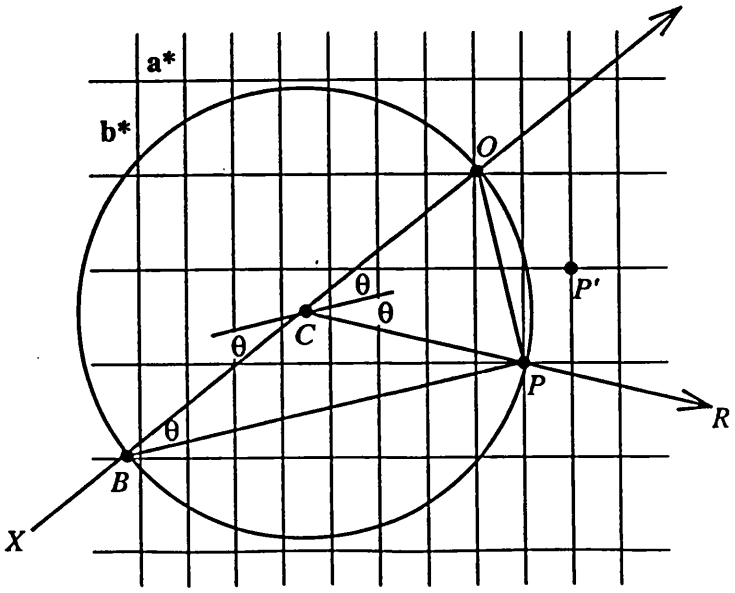


Figure 4: Reciprocal space plane a*b* showing Ewald construction [Rhodes, 1993].

Calculation of Electron density

The Structure Factor F(hkl) is the sum of the scattering by all the atoms in the unit cell and can be written for a reflection hkl as a Fourier series over a sum of n atoms in the unit cell,

$$F(hkl)=\sum_{j=1}^{n} f_j \exp[2\pi i(hx_j+ky_j+lz_j)]$$

where f_j is the scattering factor for an individual atom j (j=1,2,3....n), and x,y and z are fractional coordinates in the unit cell such that $0 \leq x,y,z \leq 1$

The structure factor can also be represented by its real and imaginary parts (figure 5).

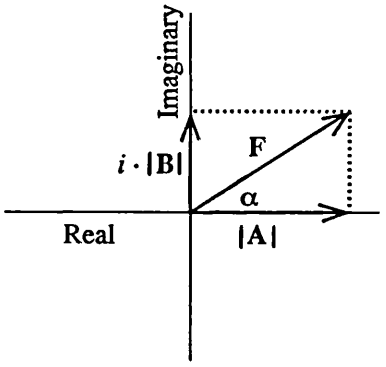


Figure 5: The Argand diagram (Rhodes, 1993)

From the Argand diagram

$$F(hkl) = |A(hkl)| + i.|B(hkl)|$$
$$= |F(hkl)|\cos\alpha(hkl) + i.\{|F(hkl)|\sin\alpha(hkl)\}$$
$$= |F(hkl)|.\exp(i\alpha)$$

where |F(hkl)| is the amplitude and α(hkl) is the phase angle.

Alternatively we can integrate over all electrons in the unit cell

$$F(hkl) = \int_V \rho(xyz)\, \exp[2\pi i(hx_j+kx_j+lz_j)]\, dxdydz$$

where V is the volume of the unit cell and $\rho(xyz)$ is the electron density at each point within the unit cell.

The Fourier Transform of the structure factor equation therefore gives the electron density equation

$$\rho\,(xyz) = 1/V \sum_{hkl} F(hkl)\, \exp[-2\pi i(hx_j+ky_j+lz_j)]$$

As the Structure Factor has a frequency, amplitude $|F(hkl)|$ and phase angle $\alpha(hkl)$ we can now write the electron density equation as

$$\rho\,(xyz) = 1/V \sum_{hkl} |F(hkl)|\, \exp[i\alpha(hkl) - 2\pi i(hx_j+ky_j+lz_j)]$$

The frequency is that of the X-ray source. From the diffraction pattern we have measured intensities I(hkl) which are proportional to the square of the amplitude of the structure factors $|F(hkl)|$ of reflection (hkl). We therefore require the value of the phase to solve the electron density equation. This is the fundamental problem in macromolecular crystallography and methods for determining phases; principly isomorphous replacement (and or anomalous scattering) and molecular replacement are discussed below.

Also, the atomic scattering factor *f* is dependent on the atomic number of the atom and must be multiplied by a temperature factor which takes into account the vibration of the atom about its mean position. For isotropic vibration

$$T(iso) = \exp[-B\sin^2\theta/\lambda^2]$$
$$= \exp[-B\{(2\sin\theta/\lambda)^2/4\}]$$
$$= \exp[-B\{(1/d)^2/4\}]$$

The thermal parameter B is related to the mean square displacement $\overline{u}^2$ of the atomic vibration, $B = (8/3)\pi^2 \times \overline{u}^2$.

Obtaining Phases

i) Isomorphous Replacement

As each atom in the unit cell contributes to each reflection in the diffraction pattern, the introduction of a 'heavy' atom (at identical sites in each unit cell) into the protein in the crystal creates a measurable difference in the observed intensities. This can be used to gain an initial estimate of the phases. This involves use of the Patterson function.

The Patterson Function, P(u) (or P(uvw)) is a Fourier summation with intensities as coefficients

$$P(uvw) = 1/V \sum_{hkl} |F(hkl)|^2 \cos[2\pi(hu+kv+lw)]$$

where u,v and w are coordinates in Patterson space. As the coefficients are squared, this means that the Patterson function is dominated by large terms. Also, as the phase angles are zero it can be calculated directly from the diffraction intensities.

P(u) can also be written as

$$P(u) = \int_{r1} \rho(r1) \times \rho(r1+u)\, dv$$

Therefore peaks in a Patterson map are at end points of vectors u centred at the origin and corresponding to vectors between all atoms in the unit cell.

The heavy atom derivative structure factor F_{PH} is the vector sum of native F_P and heavy atom F_H structure factors (figure 6). We must therefore first calculate the heavy atom positions from the isomorphous difference Patterson $(k|F_{PH}| - |F_P|)^2$ (where k is a scale factor).

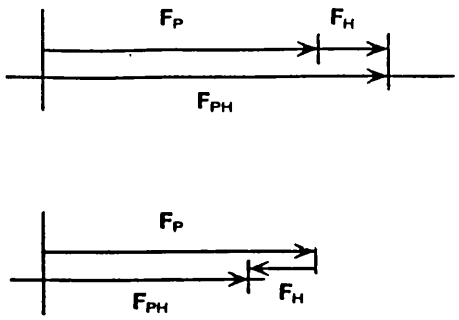


Figure 6a: Structure factors for the centric case [Drenth, 1995].

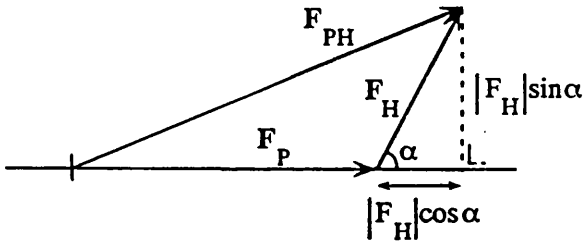


Figure 6b: Structure factors for acentric reflections ($F_{PH} = F_P + F_H$) [Drenth, 1995].

From the measured amplitudes we know the values $|F_P|$ and $|F_{PH}|$. $|F_H|$ and α_H can also be calculated from the heavy atom site coordinates. We can therefore calculate values of protein phases α_P. Except in centric cases - where α_P can have one of only two possible values- (h0l for space group $P2_1$) α_P can have any value, but a phase ambiguity arises from the fact that there are two possible solutions, symmetrical with respect to α_H and hence in general

at least two derivatives are necessary to solve the phase problem. This is shown in figure 7.

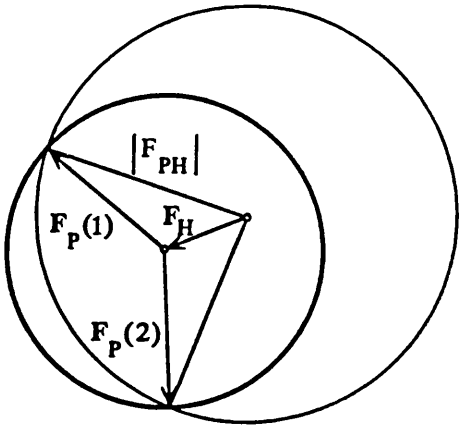


Figure 7: The Harker construction [Harker, 1956] showing phase ambiguity in the single isomorphous replacement case. Each derivative gives a possible value of the phase angle for each of the two vectors $F_P(1)$ and $F_P(2)$ [Drenth, 1995]. A second derivative should intersect the first circle at only one of these points.

There is also the consideration of experimental error in initial phase calculation, resulting from the difference Δ between $|F_{PH}|_{calc}$ and $|F_{PH}|_{obs}$. This difference is known as the 'lack of closure'.

This situation can also be described by the probability density function P(α), with the maxima of this function lying at the most probable phase value.

$$P(\alpha) = \prod_{j=1}^{N} p_j(\alpha)$$

where P_j is the probability distribution for the jth heavy atom derivative (j= 1,2,3.....n).

The reliability of the phase information for a given derivative, is also assessed by monitoring values of R_{cullis} , phasing power and FOM.

The phasing power is the ratio of the RMS heavy atom amplitude to the RMS lack of closure error.

$$\text{Phasing Power} = \left[\frac{\sum_{n} |F_H|^2}{\sum_{n} \left\{ |F_{PH}|(obs) - |F_{PH}|(calc) \right\}^2} \right]^{1/2}$$

where n is the number of oberved amplitudes for the derivative. The phasing power is a measure of the sharpness of the probability density function at the most probable phase value.

For centric reflections R_{cullis} is given by

$$R_{cullis} = \frac{\sum_{hkl} \left| |F_{PH} \pm F_P| - F_H(calc) \right|}{\sum_{hkl} |F_{PH} \pm F_P|}$$

The Figure of Merit (FOM) is the weighted mean of the cosine of deviation of the phase error (the deviation of the phase angle from α_{BEST}).

$$FOM = \frac{|F(hkl)_{BEST}|}{|F(hkl)|}$$

and

$$F(hkl)_{BEST} = \frac{\sum_{\alpha} P(\alpha) F_{hkl}(\alpha)}{\sum_{\alpha} P(\alpha)}$$

ii) Anomalous Scattering

Anomalous scattering may arise when the absorption edge of the heavy atom is near that of the X-ray wavelength, with part of the radiation being absorbed by the atom and re-emitted with a change in phase. This results in the breakdown of Friedels Law (figure 8) where |F(hkl)| is no longer equal to |F(-h-k-l)|. The anomalous scattering factor can be written in terms of its real and imaginary parts as

$$f_{ANOM} = f + \Delta f' + i\Delta f''$$

$\Delta f'$ is the dispersion component and $\Delta f''$ is the absorption component which both vary with wavelength. This information can be used in conjunction with isomorphous differences to solve the phase problem. Also, the variation of f' and f'' particularly near the absorption edge is the basis of the multiple wavelength anomalous dispersion method (MAD).

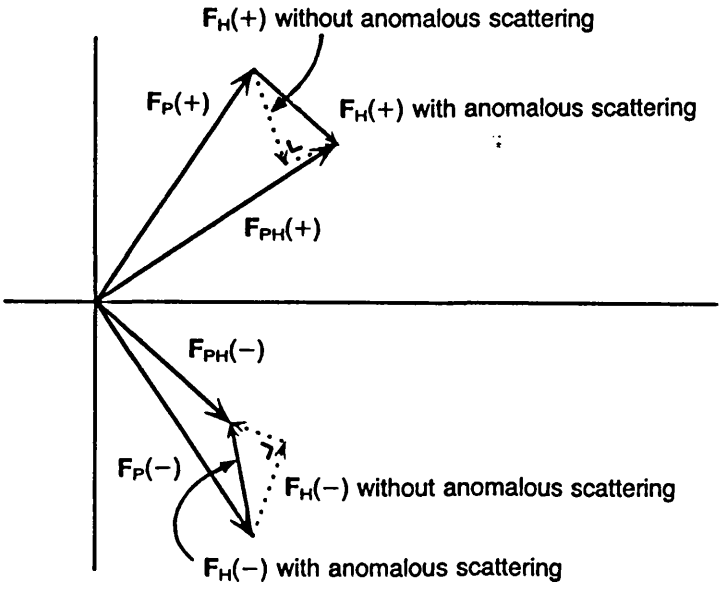


Figure 8: The effect of anomalous scattering on the structure factors for reflections (hkl) and (-h-k-l) (Friedel (bijvoet) pairs). FP(+) and FP(-) (and $F_H(+)$ and $F_H(-)$) are symmetric about the real axis in the absence of anomalous scattering [Drenth, 1993].

iii) Molecular Replacement

Phase information may also be derived for the new structure by searching for the correct orientation (determined by the rotation function) and translation (determined by the translation function) of a model (based on coordinates of a similar structure) in the new unit cell. This can be viewed in terms of Patterson space where the rotational part will involve the intra-molecular vectors which are situated within a distance from the origin equal to that of the maximum dimension the molecule. The inter-molecular vectors give information about the translational part of the problem.

The ordinary **cross rotation function** R(Ω) [Rossmann, 1972] can be defined as the overlap of the model Patterson (P_{calc}), with that of the crystal (P_{obs}) and which should have a large value on correct overlap of the two Pattersons.

$$R(\Omega) = \sum_i P_{obs}(u_i).P_{calc}(u_i, \Omega)$$

where the Pattersons vary with vector u_i and Ω is a rotation matrix.

The **self rotation function** is obtained simply by calculating the Patterson function from the X-ray data and rotating it upon itself. This can give peaks representing the non-crystallographic symmetry within an asymmetric unit.

Expanding in terms of structure factor amplitudes

$$R(\Omega) = \frac{U}{V^3} \sum_h \sum_p |F(h)|^2 |F(p)|^2 G(h,h')$$

where h and p are reciprocal space vectors and h' is a non-integral reciprocal lattice vector. $G_{h,h'}$ is an interference function which describes rotation of vector h'.

In application, many programs use Crowthers Fast rotation function [Crowther, 1972] which is calculated by expanding the Patterson density in terms of spherical harmonics rather than Cartesian Fourier components $|F(h)|^2$.

In order to identify the position of the correctly oriented model in the new unit cell we need to calculate the translation function. The model can be simply moved and structure factors calculated at each position. Again this can be viewed as a sum of observed and calculated Pattersons, based on the addition of the positional parameters, t.

$$T(\Omega,t) = \sum_i \; P_{obs}(u_i).P_{calc}(u_i, \Omega,t)$$

These can be compared with the observed structure factors and this monitored by means of a residual R-Factor (and or a correlation coefficient)

where the R-factor is calculated for a given position of the model is

$$R = \frac{\sum_{hkl} \| F(obs) | - k | F(calc) \|}{\sum_{hkl} | F(obs) |}$$

The T_2 function [Crowther & Blow, 1967] is that most commonly used in the Molecular Replacement programs, which involves a full-symmetry calculation in which all possible inter-molecular vectors in the unit cell are considered in the comparison of observed and calculated Pattersons. This is an improvement over the T_1 function for which only a pair of molecules related by the crystallographic symmetry are used.

Some of the experimental considerations involved in rotation and translation searches are discussed in chapter 4.

Refinement

Refinement of the initial model structure obtained from the above methods is necessary to gain a better agreement of the calculated structure factor amplitudes (Fc) with the observed structure factor amplitudes (Fo), scored by means of a target function. Functions commonly used include the least squares residual, the empirical energy function, or maximum likelihood.

In Least Squares refinement, the function to be minimised is the sum of the squares of differences between observed and calculated structure factors.

$$\Phi = \sum_{hkl} w(hkl)\left\{|F_{obs}(hkl)| - |F_{calc}(hkl)|\right\}^2$$

where w(hkl) is the weight typically related to confidence in a particular observation.

The function is minimised iteratively by varying positional parameters (x,y,z) and an associated isotropic temperature factor (B) for each atom in the model in an attempt to find a global minima.

Minimising a Maximum Likelihood residual, the consistency of the model is measured by the probability that given the current model, the observed values would be made. Parameters are refined by maximising the minus log likelihood L (-lnL).

If there are N observations of x_j then the likelihood is the joint probability of making the entire set of observations

$$L = p(x_1, x_2, \ldots x_N) = \prod_{j=1}^{N} p(x_j)$$

where it is commonly assumed that the observations are independent of each other.

In protein crystallography there is a relatively low ratio of observations to parameters, which can lead to poor overdetermination of the structure. Additional observations are therefore incorporated into the method; stereochemical data derived from high resolution crystal structures of small peptides can be applied [Engh and Huber, 1991] in addition to information about the non-crystallographic symmetry. The stereochemistry can be either constrained; where parameters such as bond lengths and angles are given a

fixed value to reduce the number of parameters in the model, or alternatively restrained; where the stereochemistry may vary around a standard value. Tight restraints can be used at early stages of the refinement and then loosened at a later stage.

Inputting restraints on bond lengths, the least squares target function becomes

$$\Phi = \sum_{hkl} w(hkl)\left\{|F_{obs}(hkl)| - |F_{calc}(hkl)|\right\}^2 + \sum_j W_j (d_j(\text{ideal})-d_j(\text{calc}))^2$$

wher d_j is the bond length j. This can be carried out similarly for other parameters such as bond angles, group planarities and non-bonded interactions.

The method of minimisation of the chosen target function can also vary. These methods have differing radii of convergence (distance from the energy minimum at which the method breaks down). Some methods require the gradient of the function (gradient descent). In addition to the first derivative information, full matrix methods include second derivatives (curvature).

These methods can be understood firstly by expanding the function to be minimised into a Taylor series. At the minimum (or maximum) of the function the gradient is equal to zero and therefore a shift vector between the current model and the minimum can be calculated.

In full matrix least squares minimisation there is a difficulty in calculating the transpose of the term $|d^2f(p)/dp^2|_{P=Po}$. (for the function f(p)), where Po are current parameters. Approximations can therefore be made as the diagonal and off-diagonal elements are different. The diagonal elements are affected by a single parameter, wheras the off-diagonal elements are correlated. In the sparse matrix method, pairs of off-diagonal elements which are predicted to be small are ignored. Steepest descent methods further simplify this by estimating an average value for the diagonal elements, thus increasing the radius of convergence. In the conjugate gradient method, two successive gradient calculations are compared, improving the steepest descent shift

vector. Methods such as Simulated Annealing are also often used in early stages of refinement as they have a large radius of convergence. Simulated Annealing is a molecular dynamics method (Brunger, 1992) which simulates the dynamic nature of particles. Effectively the atoms are heated to a sufficiently high temperature so that an energy barier can be overcome followed by a procees of slow cooling to allow them to reach an energy minimum. Rigid-Body refinement can also be used at an intial stage following a solution with molecular replacement to subject the model to large shifts (of the whole molecule or separate domains).

These refinement methods can be used in combination with density improvement techniques such as non-crystallographic symmetry averaging, solvent flattening and histogram matching which can help the interpretation of electron density maps when model building.

The refinement process is monitored by an R-factor, where

$$R = \frac{\sum_{hkl} \| F(obs) | - k | F(calc) \|}{\sum_{hkl} | F(obs) |} \times 100\ \%$$

In addition to this R-factor, a Free R factor [Brunger, 1992b] is also calculated using 5-10% of the data which is flagged as a test set. As the refinement is actually carried out using only the working set of relections, the Free R-factor is unbiased by the refinement process.

REFERENCES

Adams, M.W.W., Perler, F.B. & Kelly, R.M. (1995) Extremoenzymes - expanding the limits of biocatalysis. *Biotechnology* **13**, 662-668

Aguilar, C.F., Sanderson, I., Moracci, M., Ciaramella, M., Nucci, R., Rossi, M. & Pearl, L.H. (1987) Crystal structure of β-Glycosidase from the hyperthermophilic archaeon *Sulfolobus solfataricus*: resilience as a key factor in thermostability, *J. Mol. Biol.*, **271**, 789-802.

Akasako, A., Haruki, M., Oobatake, M. & Kanaya, S. (1995) High resistance of *Escherichia coli* ribonuclease HI variant with quintuple thermostabilising mutations to thermal denaturation, acid denaturation, and proteolytic degradation. *Biochemistry* **34**, 8115-8122.

Alber, T. & Mathews, B.W. (1987) Structure and Thermal stability of phage T4 lysozyme in *Methods in Enzymology*. Academic Press, London.

Alvarez, M., Zeelen, J.P., Mainfroid, V., RentierDelrue, F., Martial, J.A., Wyns, L., Wierenga, R., & Maes, D. (1998) Triose-phosphate isomerase (TIM) of the Psychrophilic bacterium *Vibrio marinus J. Biol. Chem.* **273**, 2199-2206.

Arnone, A., Bier, C.J., Cotton, F.A., Day, V.W., Hazen, E.E., Richardson, D.C., Richardson, J.S. & Yonath, A (1971) A high resolution structure of an inhibitor complex of the extracellular nuclease of *Staphylococcus aureus. J. Biol. Chem.* **246**, 2302.

Baldwin, R.L. (1986) Temperature dependence of the hydrophobic interaction in protein folding. *Proc. Nat. Acad. Sci. (USA)* **83**, 8069-8072.

Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., Pogson, C.I., Wilson, I.A., Corran, P.H., Furth, A.J., Milman, J.D., Offord, R.E., Priddle, J.D.& Waley, S.G. (1975) Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5Å resolution. *Nature* **255**, 609-614.

Barlow, D.J., & Thornton, J.M. (1983) Ion-pairs in proteins *J. Mol. Biol.*, **168**, 867.

Barton, G.J. (1990) Protein multiple sequence alignment and flexible pattern matching. *Methods in Enzymology* **183**, 403-428.

Barton, G.J. (1993) ALSCRIPT: a tool to format multiple sequence alignments *Prot. Eng.* **6**, 37-40.

Bell, G.S., Russell, R.J.M., Kohlhoff, M., Hensel, R., Danson, M.J., Hough, D.W. & Taylor, G.L. (1998) Preliminary crystallographic studies of triosephosphate isomerase (TIM) from the hyperthermophilic Archaeon *Pyrococcus woesei. Acta Cryst.* **D54**, 1419-1421.

Bernstein, B.E., & Hol, W.G.J. (1997) Probing the limits of the molecular replacement method: the case of *Trypanosoma brucei* phosphoglycerate kinase. *Acta Cryst.* **D53**, 756-764.

Bernstein, F., Koetzle, T., Williams, G., Meyer, E.J., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T., Tasumi, M. (1977) The protein data bank: a computer-based archival file for macromlecular structures. *J. Mol. Biol.* **112**, 535-542.

Blundell, T.L. & Johnson, L.N. (1976) *Protein Crystallography.* Academic Press, London.

Bogin, O., Peretz, M., Hacham, Y., Korkhin, Y., Frolow, F., Kalbgilboa, A.J. & Burstein, Y. (1998) Enhanced thermal stability of *Clostridium beijerinckii* alcohol dehydrogenase after strategic substitution of amino acid residues with prolines from the homologous thermophilic *Thermoanaerobacter brockii* alcohol dehydrogenase. *Prot. Sci.* **7**, 1156-1163.

Borchert, T.V., Abagyan, R., Kishan, K.V.R., Zeelen, J.Ph & Wieranga, R.K. (1993) The crystal structure of an engineered monomeric triosephosphate isomerase, monoTIM: the correct modelling of an eight residue loop. *Structure* **1**, 205-213.

Borchert, T.V., Kishan, K.V.R., Zeelen, J.Ph., Schliebs, W., Thanki, N., Abagyan, R., Jaenicke, R. & Wieranga, R.K. (1995) Three new crystal structures of point mutation variants of monoTIM: conformational flexibility of loop-1, loop-4 and loop-8. *Structure* **3**, 669-679.

Bradford, M.M. (1976) A rapid sensitive method for the quantitation of microgram quantities of protein utilising the principle of protein dye binding. *Anal. Biochem.* **72**, 248-252.

Bragg, W.L. (1913). *Proc. Camb. Phil. Soc.*, **17**, 43.

Branden, C., (1991) The TIM Barrel - the most frequently occurring folding motif in proteins. *Curr. Opp. Struct. Biol.* **1**, 978-983.

Bricogne, G. (1974) Geometric sources of redundancy in intensity data and their use for phase determination. *Acta Cryst.*, **A30**, 395-405.

Britton, K.L., Baker, P.J., Borges, K.M.M., Engel, P.C., Pasquo, A., Rice, D.W., Robb, F., Scandurra, R., Stillman, T.J. & Yip, K.S.P. (1995) Insights into thermal stability from a comparison of the glutamate dehydrogenase from *Pyrococcus furiousus* and *Thermococcus litoralis*. *Eur. J. Biochem.* **229**, 688-695.

Brown, J.R., & Doolittle, W.F. (1997) Archaea and the prokaryote-to-eukaryote transition. *Microbiology and Molecular Biology Reviews*, **61**, 456-502.

Brunger, A.T. (1992) X-PLOR. version 3.1. A system for X-ray crystallography and NMR. Yale University Press, New Haven.

Brunger, A.T. (1992b) The Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472-474.

Budgen, N., & Danson, M.J., (1986) Metabolism of glucose via a modified entner-doudoroff patway in the thermoacidophilic archaebacterium *Thermoplasma acidophilum*. *FEBS Lett.* **196**, 207-210

Bult, C.J., White, O., Olsen, G.J, et al (1996) Complete genome sequence of the methanogenic Archaeon Methanococcus jannaschii, *Science* **273**, 1058-1073.

Burley, S.K. & Petsko, G.A. (1985) Aromatic-aromatic interaction: a mechanism of protein structure stabilisation. *Science* **229**, 23-28.

Chan, M.K., Mukund, S., Kletzin, A., Adams, M.W.W.& Rees, D.C. (1995) Structure of a hyperthermophilic tungstopterin enzyme, aldehyde ferredoxin oxidoreductase. *Science*, **267**, 1463-1469.

Connaris, H., West, S.M., Hough, D.W., Danson, M.J. (1998) Cloning and overexpression in Escherichia coli of the gene encoding citrate synthase from the hyperthermophilic Archaeon *Sulfolobus solfataricus*. *Extremophiles*, **2**, 61-66.

Cowan, D.A. (1992) Biotechnology of the Archaea. *TIBTECH* **10**, 315-323.

Cowtan, K., (1994) Joint CCP4 and ESF-EACBM newsletter on protein crystallography **31**, 34-38.

Crick, F.H.C, & Magdoff, B. (1956) The theory of the method of isomorphous replacement for protein crystals. *Acta. Cryst.*, **9**, 901.

Crowther, R.A. & Blow, D.M. (1967) A method of positioning a known molecule in an unknown crystal structure. *Acta Cryst.* **23**. 544.

Crowther, R.A. (1972) The fast rotation function. Rossmann, M.G., (Ed), The molecular replacement method. *Int. Sci. Rev. Ser.*, **13**. 173-178. New York: Gordon & Breach.

Daggett, V. & Michael, L. (1993) Protein unfolding pathways explored through molecular dynamics simulations. *J. Mol. Biol.* **232**, 600-619.

Daniel, R.M., Dines, M., Petach, H.H. (1996) The denaturation and degradation of stable enzymes at high temperatures, *Biochem. J.* **317**, 1-11.

Danson, M.J., Hough, D.W., Russell, R.J.M., Taylor, G.L. & Pearl, L. (1996) Enzyme thermostability and thermoactivity. *Protein Eng.* **9**, 629-630.

Danson, M.J. (1988) Archaebacteria : the comparative enzymology of their central metabolic pathways. *Adv. Microb. Phys.* **29**, 166-231.

Danson, M.J. (1989) Central metabolism in the Archaebacteria - an overview. *Canadian J. Microbiol.* **35**, 58-64.

Danson, M.J., Hough, D.W. & Lunt, G.G. (Eds) (1991) The Archaebacteria: Biochemistry and Biotechnology *Biochem. Soc. Symp.* **58**

Danson , M.J. (1993) In *The Biochemistry of Archaea* Chapter 1. Elsevier Science Publishers B.V. Kates, M., Kushner, D.J., Matheson, A.T., (Eds).

Danson, M.J., Russell, R.J.M., Hough, D.W., Taylor, G.L. (1998) In Thermophiles (Wiegel, J. & Adams, M.W.W. (eds) 255-267. Comparative enzymology as an aid to understanding evolution. Taylor & Francis Ltd, London.

Davies, G., Gamblin, S.J., Littlechild, J.A. & Watson, H.C. (1993) Structure of a thermally stable 3-phosphoglycerate kinase and a comparison with its mesophilic equivalent. *Proteins* **15**, 283-289.

Dao-Pin, S., Sauer, U., Nicholson, H., & Mathews, B.W. (1991) Contributions of engineered surface salt bridges to the stability of T4 lysozyme determined by directed mutagenesis. *Biochemistry* **30**, 7142-7153.

Davenport, R.C., Bash, P.A., Seaton, B.A., Karplus, M., Petsko, G.A. & Ringe, D. (1991) Structure of the triosephosphate isomerase phosphoglycolohydroxamate complex - an analogue of the intermediate on the reaction pathway. *Biochemistry* **30**, 5821-5826.

Day, M.W., Hsu, B.T., Joshuator, L., Park, J.B., Zhou, Z.H., Adams, M.W.W. & Rees, D.C. (1992) X-ray crystal structures of the oxidised and reduced forms of the rubredoxin from the marine hyperthermophilic Archaebacterium *Pyrococcus furiousus. Protein Science* **1**, 1494-1507.

Delboni, L.F., Mande, S.C., Rentier-Delrue, F., Mainfroid, V., Turley, S., Vellieux, F.M.D., Martial, J.A. & Hol, W.G.J. (1995) Crystal structure of recombinant triosephosphate isomerase from Bacillus stearothermophilus - an analysis of potential thermostabilizing factors in 6 isomerases with known three dimensional structures points to the importance of hydrophobic interactions. *Protein Science* **4**, 2594-2604.

De Rosa, M., Gambacorta, A. & Bullock, J.D. (1975) Extremely thermophilic acidophilic bacteria convergent with *Sulfolobus acidocaldarius. Journal of General Microbiology,* **86**, 156-164.

De Rosa, M., Gambacorta, A., Nicolaus, B., Giardina, P., Poerio, E., & Buonocore, V. (1984) Glucose metabolism in the extreme thermoacidophilic archaebacterium *Sulfolobus solfataricus. Biochem. J.* **224**, 407-414.

Dill, K.A., Alonso, D.O.V., & Hutchinson, K.,(1989) Thermal stabilities of globular proteins, *Biochem.*, **28**, 5439-5449.

Dill, K.A. (1990) Dominant forces in protein folding, *Biochem.*, **29**, 7133-7155.

Dodson, E. & Baker, T. (1998) In *The CCP4 Suite- computer programs for protein crystallography* (Love, D., Ed.) p. 9 CCLRC, Daresbury.

Dodson, E. & Terry, H. (1998) In *The CCP4 Suite- computer programs for protein crystallography* (Love, D., Ed.) p. 8 CCLRC, Daresbury.

Doolittle, R.F (1995) Of Archaea and Eo -whats in a name *Proc. Natl. Acad. Sci. USA,* **92**, 2421-2423.

Drenth, J., (1995) Principles of protein X-ray crystallography. Springer, New York.

Eijsink, V.G., Vriend, G., Van der Zee, J., Van den Burg, B. & Venema, G. (1992) Increasing thermostability of the neutral proteinase of *Bacillus stearothermophilus* by improvement of internal hydrogen-bonding. *Biochem. J.* **285**, 625-628.

Elcock, A.H. (1998) The stability of salt bridges at high temperatures : Implications for hyperthermophilic proteins. *J. Mol. Biol.* **284**, 489-502.

Engh, R.A. & Huber, R. (1991) Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Cryst.* **A47**, 392-400.

Eriksson, A.E., Baase, W.A., Zhang, X.J., Heinz, D.W., Blaber, M., Baldwin, E.P. & Mathews, B.W. (1992) Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science* **255**, 178-183.

Evans, P. Dodson, E. & Dodson, R. (1998) In *The CCP4 Suite- computer programs for protein crystallography* (Love, D., Ed.) p. 27 CCLRC, Daresbury.

Ewald, P.P. (1921). *Z. Kristollagr. Miner.*, **56**, 129.

Facchiano, A.M., Colonna, G. & Ragone, R. (1998) Helix stabilising factors and stabilization of thermophilic proteins: an X-ray based study. *Prot. Eng.* **11**, 753-760.

Farber, G.K., & Petsko, G.A. (1990) The evolution of α/β barrel enzymes. *TIBS* **15**, 228-234.

Fersht, A.R., Shi, J.P., Knill-Jones, J., Lowe, D.M., Wilkinson, A.J., Blow, D.M. & Brick, P. (1985) Hydrogen bonding and biological specificity analysed by protein engineering. *Nature* **314**, 235-238.

Fersht, A.R. & Serrano, L. (1993) Principles of protein stability derived from protein engineering experiments. *Curr. Opp. Struct. Biol.* **3**, 75-83.

Forood, B., Feliciano, E.J., and Nambiar K.P. (1993) Stabilisation of α-helical structures in short peptides via end capping. *Proc.Natl.Acad.Sci. USA,* **90**, 838-842.

French, G.S., & Wilson, K.S. (1978) On the treatment of negative intensity observations. *Acta Cryst.* **A34**, 517.

Furste, J.P., Pansegrau, W., Frank, R., Blocker, H., Scholz, P., Bagdasarian, M. & Lanka, E. (1986) Molecular cloning of the plasmid RP4 primase region in a multi-host-range *tacP* expression vector. *Gene* **48**, 119-131.

Gerike, U., Danson, M.J., Russell, R.J., Hough, D.W. (1997) Sequencing and expression of the gene encoding a cold-active citrate synthase from an Antarctic bacterium, strain DS2-3R. *Eur. J. Biochem.*, **248**, 49-57.

Hanford, P.A., Ner, S.S., Bloxham, D.P., Wilton, D.C. (1988) Site-directed mutagenesis of citrate synthase: the role of the active site aspartate in the binding of Acetyl-CoA but not oxaloacetate. *Biochem. Biophys. Acta.* **953**, 232-240.

Hardy, F., Vriend, G., Veltman, O.R., van der Vinne, B., Venema, G. & Eijsink, V.G.H. (1993) Stabilisation of *Bacillus stearothermophilus* neutral protease by introduction of prolines. *FEBS Letters* **317**, 89-92.

Hatanaka, H., Tanimura, R., Katoh, S. & Inagaki, F. (1997) Solution structure of Ferredoxin from the thermophilic cyanobacterium *Synechococcus elongatus* and its thermostability. *J. Mol. Biol.* **268**, 922-933.

Hendrickson, W.A. (1985) Stereochemically restrained refinement of macromolecular structures. *Methods in Enzym.* **115**, 252-270.

Hendrickson, W.A. (1991) Determination of Macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**, 51-58.

Hendrickson, W.A. & Ogata, C.M. (1997) Phase determination from multiwavelength anomalous difraction measurements. *Methods in Enzymology* **276**, 494-523.

Hennig, M., Darimont, B., Sterner, R., Kirschner, K. & Jansonius, J.N. (1995) 2.0Å structure of indole-3-glycerol phosphate synthase from the hyperthermophile *Sulfolobus solfataricus*: possible determinants of protein stability. *Structure* **3**, 1295-1306.

Hennig, M., Sterner, R., Kirschner, K. & Jansonius, J.N. (1997) Crystal structure at 2.0Å resolution of Phosphoribosyl anthranilate isomerase from the hyperthermophile *Thermotoga maritima*: possible determinants of protein stability. *Biochemistry* **36**, 6009-6016.

Hensel, R. & Siebers, B. (1993) Glucose catabolism of the hyperthermophilic archaeon *Thermoproteus tenax*. *FEMS Microbiol. Lett.* **111**, 1-8.

Herbert, R.A. (1992) A perspective on the biotechnological potential of extremophiles. *TIBTECH* **10**, 395-401

Honig, B., & Nicholls, A. (1995) Classical electrostatics in biology and chemistry. *Science* **268**, 1144-1149.

Honig, B., Hubbell, W., & Flewelling, R.F. (1986) Electrostatic interactions in membranes and proteins. *Annu. Rev. Biophys. Chem.*, **15**, 163.

Horovitz, A., Mathews, J.M., and Fersht, A.R. (1992) α-helix stability in proteins 2. Factors that influence stability at an internal position. *J.Mol.Biol.*, **227**, 560-568.

Horovitz, A., Serrano, L., Avron, B., Bycroft, M. & Fersht, A.R. (1990) Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double mutant cycles. *Biochemistry*, **29**. 9343-9352.

Hubbard, S. J., Gross, K-H. & Argos, P. (1994) Intramolecular cavities in globular proteins. *Prot. Eng.*, **7**, 613-626.

Hutchinson, E.G. & Thornton, J.M. (1996) PROMOTIF; a program to identify and analyse structural motifs in proteins. *Prot. Sci.* **5**, 212-220.

Goldman, A. (1995) How to make my blood boil. *Structure*, **3**, 1277-1279.

Ikai, A. (1980) Thermostability and aliphatic index of globular proteins. *J. Biochem.* **88**, 1895-1898.

Imanaka, T., Shibazaki, M. & Takagi, M. (1986) A new way of enhancing the thermostability of proteases. *Nature* **324**, 695-697.

Ishikawa, K., Okumura, M., Katayanagi, K., Kimura, S., Kanaya, S., Nakamura, H. & Morikawa, K. (1993) Crystal structure of Ribonuclease H from *Thermus thermophilus* HB8 refined at 2.8 Angstrom resolution *J. Mol. Biol.* **230**, 529-542.

Jaenicke, R., Schurig, H., Beauchamp, N. & Ostendorp, R. (1996) Structure and stability of hyperstable proteins: glycolytic enzymes from hyperthermophilic bacterium *Thermatoga maritima. Adv. Protein Chem.* **48**, 181-269.

Jancarik, J. & Kim, S-H. (1991) Sparse-matrix sampling - a screening method for crystallisation of proteins. *J. Appl. Cryst.* **24**, 409-411.

John, J., Crennel, S.J., Hough, D.W., Danson, M.J. & Taylor, G.L. (1994) The crystal structure of glucose dehydrogenase from *Thermoplasma acidophilum Structure* **2**, 385-393.

Jones, S. & Thornton, J.M. (1995) Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.* **63**, 31-165.

Jones, T.A. (1978) A graphics model building and refinement system for macromolecules. *J. Appl. Cryst.* **11**, 268-272.

Joseph-McCarthy, D., Rost, L.E., Komives, E.A. & Petsko, G. A. (1994) Crystal structure of the mutant yeast triosephosphate isomerase in which the catalytic base glutamic acid 165 is changed to aspartic acid. *Biochemistry* **33**, 2824-2829.

Joseph-McCarthy, D., Lolis, E., Komives, E.A. & Petsko, G.A. (1994b) Crystal structure of the K12M/G15A triosephosphate isomerase double mutant and electrostatic analysis of the active site. *Biochemistry* **33**, 2815-2823.

Kabsch, W., & Sander, C. (1983) Dictionary of protein secondary structure-pattern recognition of hydrogen-bonded and geometrical features. *Bipolymers* **22**, 2577-2637.

Kabsch, W. (1998) In *The CCP4 Suite- computer programs for protein crystallography* (Love, D., Ed.) pp. 8 & 37, CCLRC, Daresbury.

Keeling, P.J. & Doolittle, W.F.(1995) Archaea - narrowing the gap between prokaryotes and eukaryotes. *Proc. Natl. Acad. Sci. USA* **92**, 5761-5764.

Kellis Jr., J.T., Nyberg, K., Sali, D. & Fersht, A.R. (1988) Contribution of hydrophobic interactions to protein stability. *Nature* **323**, 784-786.

Kelly, R.M. & Brown, S.H. (1993) Enzymes from high temperature microorganisms *Curr. Opp. Struct. Biotech.* **4**, 188-192

Kelly, C.A., Nishiyama, M., Ohnishi, Y., Beppu, T., & Birktoft, J.J. (1993b) Determinants of protein thermostability observed in the 1.9Å crystal structure of malate dehydrogenase from the thermophilic bacterium *Thermus flavus*. *Biochemistry*, **32**, 3913-3922.

Kimura, S., Nakamura, H., Hashimoto, T., Oobatake, M. & Kanaya, S. (1992) Stabilisation of Escherichia coli ribonuclease HI by strategic replacement of amino acid residues with those from the thermophilic counterpart. *Journal of Biological Chemistry* **267**, 21535-21542.

Kleywegt, G.J. & Jones, T.A. (1994) Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Cryst.* **D50**, 178-185.

Kleywegt, G.J. & Jones, T.A. (1996) xdlMAPMAN and xdlDATAMAN - programs for reformatting, analysis and manipulation of biomacromolecular electron density maps and reflection data sets. *Acta Cryst.* **D52**, 829-832.

Kleywegt, G.J. & Jones, T.A. (1996b) Efficient rebuilding of protein structures. *Acta Cryst.* **D52**, 829-832.

Knapp, S., de Vos, W.M., Rice, D. & Ladenstein, R. (1997) Crystal structure of glutamate deydrogenase from the hyperthermophilic Eubacterium Thermotoga maritima at 3.0Å resolution. *J. Mol. Biol.* **267**, 916-932.

Knowles, J.R. (1991) Enzyme catalysis: not different just better. *Nature* **350**, 121-124.

Kohlhoff, M.K., Dahm, A. & Hensel, R. (1996) Tetrameric triosephosphate isomerases from hyperthermophilc Archaea *FEBS Lett.* **383**, 245-250.

Komives, E.A., Chang, L.C., Lolis, E., Tilton, R.F., Petsko, G.A. & Knowles, J.R. (1991) Electrophilic catalysis in triosephosphate isomerase: the role of histidine-95. *Biochemistry* **30**, 3011-3019.

Korndorfer, I., Steipe, B., Huber, R., Tomschy, A. & Jaenicke, R. (1995) The crystal structure of holo-glyceraldehyde-3-phosphate dehydrogenase from the

hyperthermophilic bacterium *Thermotoga maritima* at 2.5Å resolution. *J. Mol. Biol.* **246**, 511-521.

Kossiakoff, A.A. (1988) Tertiary Structure is a principal determinant to Protein Deamidation, *Science* **240**, 191-194.

Kraulis, P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**, 946-950.

Kurz, L.C., Drysdale, G.R., Riley, M.C., Evans, C.T., Srere, P.A. (1992) Catalytic activity of citrate synthase: effects of amino acid changes in the acetyl-CoA binding site on transition-state analogue inhibitor complexes. *Biochem.* **31**, 7908-7914.

Laemmli, U. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680-685.

Laskowski., R. A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283-291.

Lee, C. & Levitt, C. (1991) Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature* **352**, 448-450.

Leslie, A.G.W. (1988) A reciprocal space algorithm for calculating molecular envelope using the algorithm of B.C. Wang. In CCP4 Daresbury study weekend, nos. DL/SCI/R26, ISSN 0144-5677. Warrington, UK. Daresbury Laboratory.

Liao, D-I., Karpusas, M. & Remington, S.J. (1991) Crystal structure of an open conformation of citrate synthase from chicken heart at 2.8 Angstrom resolution. *Biochem.* **30**, 6031-6036.

Lim, J-H., Yu, Y.G., Han, Y.S., Cho, S-J., Ahn, B-Y, Kim, S-H., & Cho, Y. (1997). The crystal structure of an Fe-superoxide dismutase from the hyperthermophile Aquifex pyrophilus at 1.9Å resolution : structural basis for thermostability. *J. Mol. Biol.* **270**, 259-274.

Lodi, P.J. & Knowles, J.R. (1991) Neutral imidazole is the electrophile in the reaction catalysed by triosephosphate isomerase: structural origins and catalytic implications. *Biochemistry* **30**, 6948-6956.

Lodi, P.J., Chang, L.C., Knowles, J.R. & Komives, E.A. (1994) Triosephosphate isomerase requires a positively charged active site: the role of lysine-12. *Biochemistry* **33**, 2809-2814.

Lolis, E., Alber, T., Davenport, R.C., Rose, D., Hartman, F.C. & Petsko, G.A. (1990) Structure of yeast triosephosphate isomerase at 1.9Å resolution. *Biochemistry* **29**, 6609-6618.

Lolis, E., & Petsko, G.A. (1990) Crystallographic analysis of the complex between triosephosphate isomerase and 2-phosphoglycolate at 2.5Å resolution: implications for catalysis. *Biochemistry* **29**, 6619-6625.

Luthy, R., Bowie, J.U. & Eisenberg, D. (1992) Assessment of protein models with 3-dimensional profiles. *Nature* **356**, 83-85.

Lyu, P.C., Gans, P.J., & Kallenbach, N.R. (1992) Energetic contribution of solvent-exposed ion pairs to alpha-helix structure. *J. Mol. Biol.* **223**, 343-350.

Mainfroid, V., Terpstra, P., Beauregard, M., Frere, J-M., Mande, S., Hol, W. Martial, J.A. & Goraj, K. (1996) Three hTIM mutants that provide new insights on why TIM is a dimer. *J. Mol. Biol.* **257**, 441-456.

Malakauskus, S.M. & Mayo, S.L. (1998) Design, structure and stability of a hyperthermophilic protein variant. *Nature Structure Biology* **5**, 470-475.

Mande, S.C., Mainfroid, V., Kalk, K.H., Goraj, K., Martial, J.A., Hol, W.G.J. (1994) Crystal structure of recombinant triosephosphate isomerase at 2.8 Angstrom resolution - triosephosphate isomerase-related human genetic dissorders and comparison with the trypanosomal enzyme *Protein Science* **3**, 810-821.

Mathews, B.W. (1968) Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491-497.

Mathews, B.W., Nicholson, H. & Becktel, W.J. (1987) Enhanced protein thrmostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Natl. Acad. Sci. USA.* **84**, 6663-6667.

Mathews, B.W. (1987) Genetic and structural analysis of the protein stability problem Biochemistry. **26**, 6888-6891.

Mathews, B.W. (1993) Structural and Genetic analysis of protein stability. *Annu. Rev. Biochem.* **62**, 139-160.

McDonald, I.K. & Thornton, J.M. (1994) Satisfying Hydrogen-bonding potential in proteins. *J. Mol. Biol.* **238**, 777-793.

McEvily, A.J., & Harrison, J.M. (1986) Subunit equilibria of porcine heart citrate synthase - effects of enzyme concentration, pH and substrates *J. Biol. Chem.* **261**, 2593-2598.

Meinnel, T., Lazennec, C., Dardel, F., Scmitter, J.M. & Blanquet, S. (1996) The C-terminal domain of peptide deformylase is disordered and dispensable for activity. *FEBS Letts.* **385**, 91-95.

Menendez-Arias, L. and Argos, P. (1989) Engineering protein thermostability, Sequence statistics point to residue substitutions in α-helices *J.Mol.Biol.*, **206**, 397-406.

Merkler, D.J., Farrington, G.K. & Wedler, F.C., (1981) Protein thermostability correlations between calculated macroscopic parameters and growth temperature for closely related thermophilic and mesophilic bacilli. *Int. J. Pept. Protein Res.*, **18**, 430-442.

Mrabet, N.T., Broeck, A., *et al* & Wodak, S. (1992) Arginine residues as stabilising elements in proteins. *Biochemistry* **31**, 2239-2253.

Muir, J.M., Russell, R.J.M, Hough, D.W., Danson, M.J. (1995) Citrate synthase from the hyperthermophilic Archaeon, *Pyrococcus furiousus. Prot.Eng.* **8**, 583-592.

Murshudov, G.N., Vagin, A.A., Dodson, E.J. (1997) Refinement of Macromolecular structures by the maximum-likelihood method. *Acta Cryst* **D53**, 240-255.

Navaza, J. (1994) AMoRe: an automated package for molecular replacement. *Acta. Cryst.* **A50**, 157-163

Nicholls, A., Sharp, K. & Honig, B. (1991) GRASP: graphical representation and analysis of structural properties. *Proteins* **11**, 281.

Nicholson, H., Becktel, W.J. & Mathews, B.W. (1988) Enhanced protein thermostability from designed mutations that interact with α-helix dipoles. *Nature* **336**, 651-656.

Nicholson, H., Anderson, D.E., Dao-pin, S. & Mathews, B.W. (1991) Analysis of the interaction between charged side-chains and the α-helix dipole using designed thermostable mutants of phage T4 lysozyme. *Biochemistry* **30**, 9816-9828.

Noble, M.E.M., Zeelen, J.Ph. & Wieranga, R.K. (1993) Structures of the open and closed state of trypanosomal triosephosphate isomerase, as observed in a new crystal form: implications for the reaction mechanism. *Proteins* **16**, 311-326.

Noble, M.E.M., Wieranga, R.K., Lambeir, A., Opperdoes, F.R., Thunnissen, A.W.H., Kalk, K.H., Groendijk, H. & Hol, W.G.J. (1991) The adaptability of the active site of trypanosomal triosephosphate isomerase as observed in the crystal structures of three different complexes. *Proteins* **10**, 50-69

Noble, M.E.M., Zeelen, J.Ph. & Wieranga, R.K. (1993) Structure of triosephosphate isomerase from *Escherichia coli* determined at 2.6 Angstrom resolution *Acta. Cryst.* D49, 403-417.

Olendzenski, L.& Gogarten, J.P. (1998) In Thermophiles (Wiegel, J. & Adams, M.W.W. (Eds) 165-185. Deciphering the molecular record for the early evolution of life: gene duplication and horizontal gene transfer. Taylor & Francis Ltd, London.

Olsen, G.J. & Woese, C.R. (1993) Ribosomal RNA - a key to phlogeny. *FASEB* **7**, 113-123

O'Neil, K.T. and Degrado, W.F. (1990) A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science*, **250**, 646-651.

Otwinowski, Z. & Zbyszek (1991) Maximum likelihood refinement of heavy atom parameters. In: (Wolf *et al*, 1991)

Otwinowski, Z. & Minor, W. (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307-325.

Pappenberger, G., Schurig, H. & Jaenicke, R. (1997) Disruption of an ionic network leads to accelerated thermal denaturation of D-glyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima*. *J. Mol. Biol.* **274**, 676-683.

Patton, A.J., Hough, D.W., Towner, P., & Danson, M.J. (1993) Does Escherichia coli possess a second citrate synthase gene. *Eur. J. Biochem.* **214**, 75-81.

Paupitt, R.A., Karlsson, R., Picot, D., Jenkins, J.A. Niklaus-Reimer, A-S. & Jansonius, J.N. (1988) Crystal structure of neutral protease from *Bacillus cereus* refined at 3.0Å resolution and comparison with the homologous but more thermostable enzyme thermolysin. *J. Mol. Biol.* **199**, 525-537.

Perutz, M. F. & Raidt, H. (1975) Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. *Nature* **255**, 256-259.

Pinker, R.J., Lin, L., Rose, G.D., and Kallenbach, N.R. (1993) Effects of alanine substitutions in α-helices of sperm whale myoglobin on protein stability. *Protein Science*, **2**, 1099-1105.

Priestle, J.P. (1995) Consistent stereochemical dictionaries for refinement and model building. In *CCP4 Daresbury study weekend*, nos DL-CONF-95-001, ISSN 1358-6254. Warrington, UK. Daresbury Laboratory.

Privalov, P.L., (1979) Stability of proteins. *Adv. Protein Chem.* **33**, 167.

Read, R. (1986) Improved Fourier coefficients for maps using phases from partial structure errors. *Acta Cryst.* **A42**, 140-149.

Reardon, D. & Farber, G.K. (1995) The structure and evolution of α/β-barrel proteins. *FASEB* **9**, 497-503.

Rees, D.C., & Adams, M.W.W. (1995) Hyperthermophiles: taking the heat and loving it. *Structure*, **3**, 251-254.

Remington, S., Wiegand, G., & Huber, R. (1982) Crystallographic Refinement and atomic models of two different forms of citrate synthase at 2.7 and 1.7 Angstroms resolution *J. Mol. Biol.* **158**, 111-152.

Remington, S.J. (1992a) Mechanisms of citrate synthase and related enzymes (triose phosphate isomerase and mandelate racemase). *Curr. Op. Struct. Biol.* **2**, 730-735.

Remington, S.J. (1992b) Structure and Mechanism of citrate synthase *Curr. Top. Cell. Regul.* **33**, 209-229.

Reynolds, S.J., Yates, D.W., & Pogson, C.I. (1971) Dihydroxyacetone phosphate. Its structure and reactivity with α-glycerophosphate dehydrogenase, aldolase and triose phosphate isomerase and some possible metabolic implications. *Biochem. J.* **122**, 285-297.

Rhodes, G. (1993) Crystallography made Cystal Clear. Academic Press, London.

Rice, D.W., Yip, K.S, Stillman, T.J., Britton, K.L., Fuentes, A., Connerton, I., Pasquo, A., Scandurra, R. & Engel, P.C. (1996) Insights into the molecular basis of thermal stability from the crystal structure determination of *Pyrococcus furiousus* glutamate dehydrogenase. *FEMS Microbiol. Rev.* **18**, 105-117.

Richardson, J.S. and Richardson, D.C. (1988) Amino Acid preferences for specific locations at the ends of α helices. *Science*, **240**, 1648-1652.

Robinson, A.B., and Rudd, C.J. (1974) Deamidation of glutaminyl and asparaginyl residues in peptides and proteins. *Curr. Top. Cell: Regul.* **8**, 247.

Rossmann, M. (1972) Molecular Replacement method, Gordon & Breach, New York.

Russell, R.J.M., Hough, D.W., Danson, M.J. &Taylor, G.L. (1994) The crystal structure of citrate synthase from the thermophilic archaeon, *Thermoplasma acidophilum. Structure* **2**, 1157-1167.

Russell, R.J.M., Ferguson, J.M.C., Hough, D.W., Danson, M.J. &Taylor, G.L. (1997) The crystal structure of citrate synthase from the hyperthermophilic archaeon *Pyrococcus furiousus* at 1.9Å Resolution. *Biochem.* **36**, 9983-9994.

Russell, R.J.M., Gerike, U., Danson, M.J., Hough, D.W. &Taylor, G.L. (1998) Structural adaptions of the cold-active citrate synthase from an Antarctic bacterium. *Structure* **6**, 351-361.

Sawyer, L., Isaacs, N., & Bailey, S. (eds). (1993) Data collection and processing. CCP4 Daresbury study weekend, nos. DL/SCI/R34, ISSN 0144-5677. Warrington WA4 4AD, UK: Daresbury Laboratory for Daresbury Laboratory.

Selig, M., Xavier, K.B., Santos, H., & Schonheit, P. (1997) Comparative analysis of Embden-Meyerhof and Entner-Doudoroff glycolytic pathways in hyperthermophilic archaea and the bacterium *Thermotoga. Arch. Microbiol.* **167**, 217-232.

Schafer, T., & Schonheit, P. (1991) Pyruvate metabolism of the hyperthermophilic archaebacterium *Pyrococcus furiousus.* Acetate formation from acetyl-CoA and ATP synthesis are catalysed by an acetyl-CoA synthetase (ADP-forming). *Arch. Microbiol.* **155**, 366-377.

Schafer, T. & Schonheit, P. (1993) Gluconeogenesis from pyruvate in the hyperthermophilic archaeon *Pyrococcus furiousus* : involvement in reactions of the Embden-Meyerhoff pathway. *Arch. Microbiol.* **159**, 354-363.

Scholz , S., Sonnenbichler, J., Schafer, W., & Hensel, R. (1992) Di-myo-inositol-1,1'-phosphate - a new inositol phosphate isolated from *Pyrococcus woesei. FEBS Lett.* **306**, 239-242.

Scholz, J.M., Qian, H., Robbins, V.H. & Baldwin, R.L. (1993) The energetics of ion-pair and hydrogen bonding interactions in a helical peptide. *Biochemistry* **32**, 9668-9676.

Schonheit, P., & Schafer, T., (1995) Metabolism of hyperthermophiles., *World Journal of Microbiology & Biotechnology* , **11**, 26-57.

Schurig, H., Beaucamp, N., Ostendrop, R., Jaenicke, R., Adler, E. and Knowles, J.R. (1995) Phosphoglycerate kinase and triosephosphate isomerase from the hyperthermophilic bacterium *Thermotoga maritima* form a covalent bifunctional enzyme complex. *EMBO J.* **14**, 442-451.

Serrano, L., Sancho, J., Hirshberg, M., and Fersht, A.R. (1992) α-helix stability in proteins 1. Empirical Correlations Concerning Substitution of Side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. *J.Mol.Biol.*, **227**, 544-559.

Sheldrick, G.M. (1991) Heavy atom location using SHELXS-90. In: (Wolf *et al*, 1991)

Shirley, B.A., Stanssens, P., Hahn, U. & Pace, C.N. (1992) Contribution of hydrogen bonding to the conformational stability of ribonuclease T1 *Biochemistry* **31**, 725-732.

Skrzynski, T. (1998) In *The CCP4 Suite- computer programs for protein crystallography* (Love, D., Ed.) pp. 8 & 37, CCLRC, Daresbury

Soltis, S.M., Stowell, M.H.B, Wiener, M.C., Phillips, G.N. & Rees, D.C. (1997) Successful flash-cooling of Xenon derivatized myoglobin crystals. *J. Appl. Cryst.* **30**, 190-194.

Spassov, V.Z. Karshikoff, A.D. & Ladenstein, R. (1995). The optimisation of protein-solvent interactions : thermostability and the role of hydrophobic and electrostatic interactions. *Protein Sci.* **4**, 1516-1527.

Starich, M.R., Sandman, K., Reeve, J.N., & Summers, M.F. (1996) NMR structure of HMfB from the hyperthermophile, *Methanothermus fervidus* confirms that this archaeal protein is a histone. *J. Mol. Biol.* **255**, 187-203.

Sutcliffe, *et al* (1987) *Prot. Eng.* **1**, 337-384.

Tanner, J.J., Hecht, R.M., & Krause, K.L. (1996) Determinants of enzyme thermostability observed in molecular structure of *Thermus aquaticus* D-glyceraldehyde-3-phosphate dehydrogenase at 2.5Å resolution. *Biochem.*, **35**, 2597-2609.

Taylor, G.L. & Russell, R.J.M. (1995) Engineering thermostability - lessons from thermophilic proteins. *Curr. Opp. Struct. Biotech.* **6**, 370-374.

Tickle, I., (1985) In *Proceedings of the Daresbury Study Weekend* (Machin, P.A., Ed.)., p. 22-26, SERC, Daresbury.

Tickle, I. (1998) In *The CCP4 Suite- computer programs for protein crystallography* (Love, D., Ed.) p. 8 CCLRC, Daresbury.

Ten Eyck, L.F. (1998) In *The CCP4 Suite- computer programs for protein crystallography* (Love, D., Ed.) p. 9 CCLRC, Daresbury.

Teplyakov, A., etal (1990) The crystal structure of thermitase at 1.4Å resolution. *J. Mol. Biol.* **214**, 261-279.

Tomlinson, G.A., Koch, T.K., & Hochstein, L.I. (1974) The metabolism of carbohydrates by extremely halophilic bacteria: glucose metabolism via a modified Entner-Doudoroff pathway. *Can. J. Microbiol.* **20**, 1085-1091.

Tong, L. (1993) Replace: a suite of computer programs for molecular replacement calculations, *J. Appl. Cryst.* **26**, 748-751.

Tong, L., & Rossmann, M.G. (1990) The locked rotation function. *Acta Cryst.* **A46**, 783-792.

Trentham, D.R., McMurray, C.H. & Pogson, C.I. (1969) The active chemical state of D-Glyceraldehyde 3-phosphate in its reactions with D-glyceraldehyde 3-phosphate dehydrogenase, aldolase and triose phosphate isomerase. *Biochem. J.* **114**, 19-24.

Tronrud, D.E. (1994) Methods of minimization and their implications In *CCP4 Daresbury study weekend*, nos DL/SCI/R35, ISSN 0144-5677. Warrington, UK. Daresbury Laboratory.

Tronrud, D.E. (1997) TNT refinement package. *Methods in Enzym.* **277**, 306-319.

Van den Burg, B., Vriend, G., Veltman, O.R., Venema, G. & Eijsink, V.G. (1998) Engineering an enzyme to resist boiling. *Proc. Nat. Acad. Sci. USA* **95**, 2056-2060.

Varghese, J.N., Laver, W.G. & Colman, P.M. (1983) Structure of the infulenza virus glycoprotein antigen neuraminadase at 2.9 Angstrom resolution. *Nature* **303**, 35-40.

Velankar, S.S., Ray, R.S., Gokhale, S., Suma, H., Balaram, P., Balaram, M., & Murthy, N. (1997) Crystal structure of recombinat triosephosphate isomerase from *Plasmodium falciparum* at 2.2 angstrom resolution: implications for drug design *Structure* **6**, 751-761

Vieille, C. & Zeikus, J.G. (1996) Thermoenzymes : identifying molecular determinants of protein structural and functional stability. *Trends Biotechnol.* **14**, 183-190.

Viguera, A.R., Villegas, V., Aviles, F.X. & Serrano, L. (1996) Stabilisation of proteins by rational design of alpha-helix stability using helix/coil transition theory. *Folding and design* **2**, 23-33.

Villeret, V., Clantin, B., Tricot, C., Legrain, C., Roovers, M., Stalon, V., Glansdorff, N. & Van Beeumen, J. (1998) The crystal structure of *Pyrococcus furiuosus* ornithine carbamoyltransferase reveals a key role for oligomerisation in enzyme stability at high temperatures. *Proc. Nat. Acad. Sci. USA*, **95**, 2801-2806.

Vogt, G., Woell, S., & Argos, P. (1997) Protein thermal stability, hydrogen bonds and ion pairs. *J. Mol. Biol.* **269**, 631-643.

Wade, R.C. (1996) Brownian dynamics simulations of enzyme-substrate encounter. *Biochem. Soc. Trans.* **24**, 254-259.

Waley, S.G. (1973) Refolding of triosephosphate isomerase. *Biochem. J.* **135**, 165-172.

Walker, J., Wonacott, A.J. & Harris, J.I. (1980) Heat stability of a tetrameric enzyme, D-gylceraldehyde dehydrogenase. *Eur. J. Biochem.* **108**, 581-586.

Wallon, G., Kryger,G., Lovett, S.T., Oshima, T., Ringe, D., & Petsko, G.A. (1997) Crystal structures of *Escherichia coli* and *Salmonella typhimurium* 3-isopropylmalate dehydrogenase and comparison with their thermophilic counterpart from *Thermus thermophilus*. *J. Mol. Biol.* **266**, 1016-1031.

Wang, Bi-Cheng. (1985) Resolution of phase ambiguity in macromolecular crystallography. *In* Wycoff *et al*, 1985.

Watanabe, K., Chishiro, K., Kitamura, K. & Suzuki, Y. (1991) Proline residues responsible for thermostability occur with high frequency in the loop regions of an extremely thermostable oligo-1,6-glucosidase from *Bacillus thermoglucosidasius* KP1006. *J. Biol. Chem.* **266**, 24287-24294.

Watanabe, K., Masuda, T., Ohashi, H., Mihara, H. & Suzuki, Y. (1994) Multiple proline substitutions cumulatively thermostabilise *Bacillus cereus* ATCC7064 oligo-1,6-glucosidase. Irrefragable proof supporting the proline rule. *Eur. J. Biochem.* **226**, 277-283.

Watanabe, K., Hata., Y., Kizaki, H., Katsube, Y. & Suzuki, Y., (1997) The refined crystal structure of *Bacillus cereus* oligo-1,6-glucosidase. Structural characterisation of proline substitution sites for protein thermostabilisation. *J. Mol. Biol.* **269**, 142-153.

Weitzman, P.D.J., & Danson, M.J. (1976) Citrate synthase *Curr. Top. Cell. Regul.* **10**, 161-204.

Wierenga, R.K., Kalk, K.H. & Hol, W.G.J. (1987) Structure determination of the glycosomal triosephosphate isomerase from *Trypanosoma brucei brucei* at 2.4Å resolution. *J. Mol. Biol.* **198**, 109-121.

Wierenga, R.K., Noble, M.E.M., Postma, J.P.M., Groendijk, H., Kalk, K.H., Hol, W.G.J. & Opperdoes, F.R. (1991) The crystal structure of the open and closed conformation of the flexible loop of trypanosomal triosephosphate isomerse. *Proteins* **10**, 33-49.

Wierenga, R.K., Borchert, T.V. & Noble, M.E.M. (1992) Crystallographic binding studies with triosephosphate isomerase: conformational changes induced by substrate and substrate analogues *FEBS Letters.* **307**, 34-39.

Woese, C.R. & Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA* **74**, 5088-5090.

Woese, C.R. (1987) Bacterial Evolution. *Microbiol. Rev.* **51**, 221-271.

Woese, C.R., Kandler, O. & Wheelis, M.L. (1990) Towards a natural system of organisms - proposal for the domains Archaea, Bacteria and Eukarya. *Proc. Natl. Acad. Sci. USA* **87**, 4576-4579.

Woese, C.R. In *The Biochemistry of Archaea* Introduction (1993) Elsevier Science Publishers. B.V. Kates, M., Kushner, D.J., Matheson, A.T., (Eds).

Wolf, W., Evans, P., & Leslie, A.G.W (eds) (1981). Isomorphous replacement and anomalous scattering. CCP4 Daresbury study weekend, nos. DL/SCI/R32, ISSN 0144-5677. Warrington, UK. Daresbury Laboratory.

Wycoff, H., Hirs, C.H.W. & Timasheff, S.N. (eds) (1985) Diffraction methods for biological macromolecules. *Methods in Enzym.* **115**. Academic Press.

Yip, K.S.P., Stillman, T.J., Britton, K.L., Artymuik, P.J., Baker, P.J., Sedelnikova, S.E., Engel, P.C., Pasquo, A., Chiaraluce, R., Consalvi, V., Scandurra, R. & Rice, D.W. (1995) The structure of *Pyrococcus furiousus* glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures. *Structure* **3**, 1147-1158.

Zhang, Z., Sugio, S., Komives, E.A., Liu, K.D., Knowles, J.R., Petsko, G.A., & Ringe, D. (1994) Crystal structure of recombinant chicken triosephosphate isomerase phosphoglycolohydroxamate compex at 1.8 Angstrom resolution. *Biochem.* **33**, 2830-2837.

Zhang, K.Y.J. & Main, P. (1990) The use of the Sayre equation with sovent flattening and histogram matching for phase extension and refinement of protein structures. *Acta Cryst.*, **A46**, 377-381.

Zhou, H., Wong, K. & Vijayakumar, M. (1997) Design of fast enzymes by optimising interaction potential in the active site. *Proc. Nat. Acad. Sci. USA* **94**, 12372-12377.

Zillig, W., Ingelore, H., Klenk, H, Trent, J., Wunderl, S., Janekovic, D., Imsel, E. & Haas, B. (1987) *Pyrococcus woesei*, sp. nov., an ultra-thermophilic marine archaebacterium representing a novel order, *Thermococcales. System. Appl. Microbiol.* **9**, 62-70.

Zuber, H. (1988) Temperature adaption of lactate dehydrogenase: Structural functional and genetic aspects. *Biophys. Chem.* **29**, 171-179.